Part 3- Session Papers for the EPA 23rd Annual National Conference on Managing Environmental Quality Systems

April 13-16, 2004 Tampa, Florida

Analytical Software for Environmental Analysis

Review of Asymmetric Confidence Intervals and Evaluation of "WesVar" Software for Analysis of Data from NHANES Complex Surveys (Hans Allender, U.S. EPA)

Cost Effective "Collaborative Sampling" in Visual Sample Plan (VSP) Software To Estimate Means and Test Hypotheses (Richard O. Gilbert, Pacific Northwest National Laboratory)

The Use of F/S plus Geostatistical Module (John Bing-Canar, U.S. EPA)

Innovative Statistical Methodologies

Determining Detection Limits for Environmental Analyses (Thomas Georgian, U.S. Army)

A Statistical Methodology for Estimating Background Concentrations (Basil Coutant, Battelle)

Overdispersion Models for the Violation of Nitrate Concentration Limits (Nagaraj Neerchal, University of Maryland)

Assessing Environmental Health Statistics

Measuring Pesticides, Lead, Allergens, and other Dangers in Homes (John Rogers, Westat)

Pesticide Epidemiology, Biomonitoring, and Risk Assessment: Four Case Examples (Ruth Allen, U.S. EPA)

Combining Environmental Indicators (Bimal Sinha, University of Maryland)

Environmental Statistical Modeling

Modeling Hazard Waste Arrival and Single Server Incinerator in Fixed Time: A Monte-Carlo Approach (Nelson Andrews, U.S. EPA)

Region 5 Changes in Estimated Hazard Exposure and Demographic Characteristics: 1990 to 2000 (Larry Lehrman and Arthur Lubin, U.S. EPA)

Review of Asymmetric Confidence Intervals and Evaluation of "WesVar" Software for Analysis of Data from NHANES Complex Surveys

Dr. Hans D. Allender, U.S. EPA

Introduction

Within the continuous expansion of statistical capabilities at USEPA Office of Pesticide Programs (OPP) and through the introduction to the "National Health and Nutrition Examination Survey" (NHANES) database and results, the NHANES team has been exposed to complex survey designs. Many questions related to statistical procedures that apply to those complex bias designs have been presented. At the crux of the matter is the need for the NHANES to make sure that certain subgroups in the population are properly represented in the survey. To assure proper representation it is necessary to over sample these subgroups. The result of over-sampling implies that samples are not purely collected at random and in order to compensate for the bias introduced by the sample design it is required to balance the survey. The balance is reached by given a representative weight to each element in the sample. Through a complex mathematical procedure involving the Jackknife design option, NHANES estimates the proportion of subjects, and weights are assigned to each subject. In the process fifty-two replicates are created.

A normal practice to calculate point estimates and its confidence intervals from survey data is to use large-sample normal approximations. In these cases, for example, a 95% confidence intervals on a point estimated of, say, a given percentiles are often computed by adding and subtracting from the point estimate a quantity equal to twice its standard error. This normal approximation method may not be adequate, when estimating the proportion of subjects above or below a selected value (especially when the proportion is near 0.0 or 1.0 or when the effective sample size is small). In addition, confidence intervals on proportions deviating from 0.5 are not theoretically expected to be symmetric around the point estimate. Also, adding and subtracting a multiple of the standard error to an estimate near 0.0 or 1.0 can lead to impossible confidence limits (i.e., proportion estimates below 0.0 or above 1.0).

On the "Second National Report on Human Exposure to Environmental Chemicals" published by the Department of Health and Human Services, Center for Disease Control and Prevention there are a series of tables describing the Geometric mean and selected Percentiles of concentrations in urine on a variety of the compound surveyed by NHANES. A characteristic of these point estimates is a 95% confidence interval associated with each estimate. A closer look at these confidence intervals shows that they are not symmetric; they are not calculated by classical statistical methods. A brief reference in the publication describes a cumbersome procedure to obtain these intervals.

Objectives

The main goal of this effort is to find an alternative way to calculate the asymmetric 95% confidence intervals presented in the "Second National Report on Human Exposure to Environmental Chemicals". Also, to develop capabilities to operate non-classical statistical methods; especially methods of replication, this is, the creation of a great amount of sub-samples from the whole sample that will produce replicate estimates. In developing these capabilities the software WesVar version 4.2 was used and evaluated.

Methodology

Non-classical Statistical methodologies have developed because of the advances of computers that allow the "brute-force-approach" of advance processors to generate a great amount of sub-samples from the original one; and through special algorithms calculate different parameters. Because the increased precision of these methodologies, confidence intervals that with classical statistics were calculated by adding and subtracting a fix amount (x times the standard error), and so rendered symmetric, are now computer generated resulting in asymmetric intervals.

The typical distributions presented in NHANES' concentration of pesticides in urine are well skew to the right and as a reflection of this fact asymmetric confidence intervals calculated with non-classical statistics will also have the tendency to be skew to the right. In the "Second National Report on Human Exposure to Environmental Chemicals" tables for a series of compound's concentration in urine are presented showing the Geometric Mean and selected percentiles (10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}), all these point estimates are provided with asymmetric 95% confidence interval.

To independently obtain results similar to the published tables, the analyst picked Table 171, page 206 of the aforementioned report as a prototype. This table describes the concentration of 1-Naphthol; a metabolite of Carbaryl, hence, the two names will be used interchangeable. The Table is reproduced here as Table No. 1. The complete report can be found at http://www.cdc.gov/exposurereport/

The first step in the process is from the raw data to prepare the distribution of concentration of Carbaryl and to create the divisions in the data as they relate to the table we want to reproduce i.e.; Age group, Gender, and Race/ethnicity. This preparation was done with the help of the statistical package JMP 5.0.1 and using the latest database provided by NHANES 99+. Once the data was arranged in a proper way, the second step was to convert these files to files compatible with the WesVar software. Corresponding formulas were introduced in WesVar to calculate the Geometric mean and selected percentiles. Also, replicates already developed by NHANES were used. In the calculation the variance estimation method used was Jackknife n (JKn), which is the design used when the number of Primary Sampling Units (PSUs) in a stratum is greater than or equal to two, as in the case of NHANES.

The results of each program run are listed on Table No. 2 to Table No. 10. Each table represents a line in the NHANES table

Table No. 1.- Table 1-Naphthol (Carbaryl) from the Second National Report on Human Exposure to Environmental Chemicals

Table 171. 1-Naphthol

Geometric mean and selected percentiles of urine concentrations (in µg/L) for the U.S. population aged 6 to 59 years, National Health and Nutrition Examination Survey, 1999-2000.

	Geometric mean (95% conf. interval)	Selected percentiles (95% confidence interval)						Sample size
		10th	25th	50th	75th	90th	95th	
Total, age 6-59	1.70 (1.38-2.09)	< LOD	< LOD	1.22 (1.00-1.60)	2.72 (1.90-3.76)	6.20 (4.10-9.60)	12.0 (7.20-19.0)	1998
Age group								
6-11 years	*	< LOD	< LOD	1.11 (<LOD-1.60)	2.30 (1.50-3.10)	3.61 (3.00-5.10)	5.60 (4.20-11.0)	483
12-19 years	1.54 (1.22-1.94)	< LOD	< LOD	1.20 (<LOD-1.50)	2.16 (1.60-3.80)	6.00 (3.20-11.0)	8.70 (5.20-19.0)	682
20-59 years	1.79 (1.43-2.23)	< LOD	< LOD	1.40 (1.10-1.70)	2.90 (2.10-4.10)	6.60 (4.20-12.0)	14.0 (7.20-22.0)	833
Gender								
Males	1.73 (1.42-2.11)	< LOD	< LOD	1.40 (1.10-1.80)	2.90 (2.00-3.90)	6.60 (4.40-9.00)	11.0 (7.20-16.0)	974
Females	1.67 (1.33-2.10)	< LOD	< LOD	1.30 (<LOD-1.63)	2.60 (1.80-3.80)	6.20 (3.71-13.0)	14.0 (6.30-22.0)	1024
Race/ethnicity								
Mexican Americans	1.48 (1.19-1.85)	< LOD	< LOD	1.10 (<LOD-1.70)	2.20 (1.60-3.10)	4.50 (3.00-6.70)	7.70 (4.60-14.0)	697
Non-Hispanic blacks	1.80 (1.42-2.28)	< LOD	< LOD	1.38 (1.10-1.80)	3.10 (1.89-4.80)	7.10 (4.40-13.0)	13.0 (7.20-41.0)	524
Non-Hispanic whites	1.70 (1.32-2.18)	< LOD	< LOD	1.30 (<LOD-1.60)	2.70 (1.80-3.90)	5.90 (3.70-11.0)	11.0 (5.90-21.0)	602

< LOD means less than the limit of detection, which is 1.0 µg/L.

* Not calculated. Proportion of results below limit of detection was too high to provide a valid result.

Table No. 2 Total Carbaryl (URXCAR)

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
Geo	1.70	0.177	1.34	2.05	1998
Q10	0.71 .	. .			1998
Q25	0.71 .	. .			1998
Q50	1.30	0.154	1.07	1.69	1998
Q75	2.73	0.454	2.00	3.82	1998
Q90	6.25	1.447	4.29	10.09	1998
Q95	11.63	3.336	8.07	21.46	1998

Table No. 3 Carbaryl for Age 6 to 11

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoA	1.41	0.148	1.11	1.71	483
Q10A	0.71	.	.	.	483
Q25A	0.71	.	.	.	483
Q50A	1.11	0.219	0.71	1.59	483
Q75A	2.26	0.405	1.50	3.13	483
Q90A	3.61	0.604	3.10	5.53	483
Q95A	5.53	1.600	4.46	10.88	483

Table No. 4 Carbaryl for Age 12 to 19

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoB	1.54	0.177	1.19	1.90	682
Q10B	0.71	.	.	.	682
Q25B	0.71	.	.	.	682
Q50B	1.15	0.198	0.74	1.53	682
Q75B	2.14	0.601	1.70	4.11	682
Q90B	5.93	1.887	3.69	11.26	682
Q95B	9.14	6.234	6.13	31.15	682

Table No. 5 Carbaryl for Older than 20

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoC	1.79	0.197	1.39	2.18	833
Q10C	0.71	.	.	.	833
Q25C	0.71	.	.	.	833
Q50C	1.40	0.176	1.08	1.79	833
Q75C	2.92	0.471	2.19	4.08	833
Q90C	6.78	2.176	4.50	13.23	833
Q95C	13.41	3.972	8.45	24.39	833

Table No. 6 Gender: Males

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoM	1.73	0.172	1.38	2.07	974
Q10M	0.71	.	.	.	974
Q25M	0.71	.	.	.	974
Q50M	1.38	0.175	1.09	1.79	974
Q75M	2.85	0.507	2.06	4.09	974
Q90M	6.53	1.246	4.70	9.70	974
Q95M	10.41	2.362	8.16	17.64	974

Table No. 7 Gender: Females

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoF	1.67	0.191	1.29	2.05	1024
Q10F	0.71	.	.	.	1024
Q25F	0.71	.	.	.	1024
Q50F	1.27	0.227	0.78	1.69	1024
Q75F	2.62	0.480	1.90	3.82	1024
Q90F	6.17	2.395	3.93	13.54	1024
Q95F	13.35	5.534	7.63	29.84	1024

Table No. 8 Mexican Americans

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoMex	1.48	0.163	1.15	1.81	697
Q10Mex	0.71				697
Q25Mex	0.71		.	.	697
Q50Mex	1.13	0.252	0.71	1.72	697
Q75Mex	2.27	0.374	1.79	3.29	697
Q90Mex	4.34	1.112	3.11	7.58	697
Q95Mex	7.86	2.180	5.62	14.37	697

Table No. 9 Blacks

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoBla	1.81	0.215	1.38	2.24	508
Q10Bla	0.71	.	.	.	508
Q25Bla	0.71	.	.	.	508
Q50Bla	1.39	0.186	1.04	1.78	508
Q75Bla	2.98	0.705	2.06	4.89	508
Q90Bla	7.18	3.154	4.69	17.35	508
Q95Bla	12.96	10.771	8.78	52.01	508

Table No. 10 Whites

STATISTIC	ESTIMATE	STDERROR	LOWER 95%	UPPER 95%	N
GeoW	1.70	0.212	1.28	2.13	587
Q10W	0.71	.	.	.	587
Q25W	0.71	.	.	.	587
Q50W	1.31	0.183	1.06	1.79	587
Q75W	2.74	0.533	1.90	4.04	587
Q90W	6.20	2.087	4.03	12.40	587
Q95W	11.42	4.606	6.59	25.08	587

Result Analysis

As expected, when comparing the results of WesVar with the NHANES table the results are not exactly the same, however, they are remarkably closed and in many cases the same. For example, the value of the total distribution of Carbaryl calculated by NHANES for the Geometric mean is 1.70 and its 95% confidence interval is 1.38 to 2.09; the WesVar estimated is 1.70 and its 95% confidence interval 1.34 to 2.05. In this case, WesVar produces the same point estimated for the Geometric mean and moves the interval a bit to the left keeping the same proportion. Because of increase in the distribution's instability at the extremes, more acute differences could be expected in the very extreme estimates i.e.; the 95th percentile; the NHANES estimated is 12.0 with 95% CI 7.20 to 19.0 while the WesVar estimated is 11.63 with 95% CI 8.07 to 21.46. Observe that in this case WesVar produces point estimated very close to the NHANES' and shifts the interval a bit to the right.

The other point estimates of NHANES are very close to the point estimates by WesVar, as well as the confidence interval. A small discrepancy in the sample size for Blacks and Whites is observed in the WesVar results, this is because the latest database used by the analyst has some corrections that include migration of individuals to the "Other Races" group. However, point estimates on these groups are about the same.

Summary

The duplication of the Carbaryl values on Table No.1 shows that WesVar has excellent capabilities and through a much simple procedure than the one used by NHANES very reliable estimates and confidence intervals can be obtained.

Evaluation of the "WesVar" software showed several special features including

- Easy-to-use Window interface
- Data management and flexibility to use popular statistical formats files (SAS transport, ASCII, Excel, Access, etc.)
- Capabilities for weight creation when not provided by the original survey
- Several replication methods (JK1, JK2, JKn BRR, Fay, etc)
- Results in form of tables with means, totals, percentages, standardized rates, different tests, etc.
- Regression for lineal, dichotomous and multinomial logistic models

Cost Effective "Collaborative Sampling" in Visual Sample Plan (VSP) Software to Estimate Means and Test Hypotheses

Richard O. Gilbert, Brent A. Pulsipher and John E. Wilson
Pacific Northwest National Laboratory
Richland, Washington

Selecting a cost-efficient sampling design for determining the right type, number and location of environmental samples is a critically important component of any environmental study. This paper presents an innovative design called Collaborative Sampling (CS) that can be more cost-effective in some situations than simple random sampling. The CS design uses two measurement methods: a field-based relatively inexpensive measurement method, and the standard laboratory "expensive" method. The idea behind CS is to replace the need for obtaining so many expensive measurements with collecting a larger number of the less expensive measurements. The CS design is currently being added to the suite of designs in the Visual Sample Plan (VSP) software, which can be downloaded free at http://dqo.pnl.gov/vsp. This paper discusses the CS methodology, assumptions and VSP implementation of the CS design.

1.0 Introduction

The importance of selecting a sampling design for obtaining representative environmental data for decision making cannot be disputed. The application, benefits, and limitations of several basic and innovative sampling designs are discussed in EPA (2002). The Collaborative Sampling (CS) design, which is not discussed in EPA (2002), can be more cost effective in some situations than simple random sampling for estimating the mean and testing hypotheses about the mean. Although CS may be new to many environmental professionals, discussions of the CS design can be found in, e.g., Gilbert (1987) under the title of "Double Sampling."

The CS design uses two measurement methods: the "standard analysis" (sometimes called the laboratory analysis or "the expensive method") and a less expensive and possibly less accurate measurement method (sometimes called the field-based analysis or "the inexpensive method"). The idea behind CS is to replace the need for obtaining so many expensive measurements with collecting a larger number of the less expensive measurements. The inexpensive method is used at n' locations and the expensive method is used at n of those n' locations, where n' is typically much larger than n.

The CS design is currently being implemented in the Visual Sample Plan (VSP) software tool for use when the sampling objective is to estimate a mean, compute an upper confidence limit on the mean, or test whether the mean exceeds an upper threshold value. VSP is map-based, user-friendly visual tool that helps the user determine the number and location of samples needed to ensure confident decisions. It is focused primarily on sampling design but some modules, including the CS module, incorporate statistical analysis routines for analyzing the data once it has been gathered.

2.0 Estimating the Mean

Suppose the objective of sampling is to estimate the mean of a contaminant in surface soil over a defined geographical region. One design that might be considered is simple random sampling (or perhaps systematic grid sampling) to select sampling locations, and then use the standard ("expensive") laboratory analysis method on the collected samples. Should the CS design be used instead? As discussed in Gilbert (1987, Chapter 9), the following conditions must hold for CS to be more cost effective than using the entire measurement budget to obtain expensive measurements on samples collected using a simple random sampling design:

- There is an underlying linear regression relationship between the two types of measurements
- There is a sufficiently high correlation, ρ, between the two types of measurements made at the same locations
- The ratio $R = C_{ex} / C_{inex}$ is sufficiently large, where C_{ex} is the cost of a single expensive measurement and C_{inex} is the cost of a single inexpensive measurement.

When the objective is to estimate the mean, CS will be more cost efficient than simple random sampling if the following inequality holds (Gilbert 1987, equation 9.5):

$$\rho^2 > \frac{4R}{(1+R)^2} \qquad (1)$$

In practice, the value of ρ will be uncertain and should be estimated using a "pilot" study in which the proposed inexpensive and expensive measurement methods are used in realistic field and laboratory conditions for, say 20 or more locations. Also, these pilot study data should be plotted in a regression scatter plot to assess the linearity assumption.

If CS is cost effective, then equations in Gilbert (1987, page 109) can be used to compute the number of samples, n' and n, needed. Gilbert provides equations for two cases:

- Minimize the variance of the estimated mean for a given fixed measurement budget
- Minimize the total measurement cost subject to the constraint that the variance of the estimated mean is no greater than the variance of the mean that would be obtained based on n expensive measurements obtained using a simple random sample design.

The above methodology (testing for cost efficiency and computing n' and n when the sampling objective is to estimate the mean) can be easily accomplished using the VSP software code. After booting up VSP, simply click on **Sampling Goals > Estimate the Mean > Data Not Required to be Normally Distributed > Collaborative Sampling > Simple Random Sampling** or **Systematic Grid Sampling** to access the dialog box for inputting the required Data Quality Objectives (DQOs).

3.0 Confidence Limits on the Mean

Suppose the sampling objective is to estimate the mean and also compute a one-sided upper or lower confidence limit or a two-sided confidence interval on the mean. In additional to providing an interval within which there is confidence the true mean lies, an upper confidence limit on the mean is sometimes used to test if the mean exceeds a threshold value. A method for computing the required n' and n samples when the objective is to compute confidence limits has recently been developed by the authors and is currently being incorporated into VSP. The VSP user gains access to this method in VSP by clicking **Sampling Goals > Construct Confidence Interval on the Mean > Can Assume Data will be Normally Distributed > Collaborative Sampling > Simple Random Sampling** or **Systematic Grid Sampling**.

This CS module works very similarly to the CS VSP module discussed in Section 2.0 above. First, the VSP user inputs the following DQOs into the VSP dialog box (the desired width of the confidence interval, the desired confidence level, the expected total standard deviation of the data set of expensive measurements, the expected correlation ρ between the inexpensive and expensive measurements, and costs C_{ex} and C_{inex}). Then VSP determines if CS is cost effective using Equation 1.0 above.

If CS is cost effective, then VSP computes n' and n such that the total measurement cost, C, is minimized subject to the constraint that the width of the confidence interval (CI) will be no greater than a CI width that would be obtained using n_v samples obtained using simple random sampling and measured using only the expensive measurement method. This value of n_v is computed using an iterative procedure (Gilbert 1987, page 30). Then n' and n are computed using n_v and Equations 9.8, 9.9 and 9.10 in Gilbert (1987, page 109).

After the n' and n measurements have been obtained, the VSP user can enter them into VSP. Then VSP computes:

- the mean, $\bar{y}_{cs}$, and it's standard deviation, $s_{\bar{y}_{cs}}$, using Equations 9.1 and 9.2, respectively, in Gilbert (1987, page 107)
- the confidence interval on the mean assuming the data are normally distributed or that n' and n are large enough such that the estimated mean is normally distributed
- the estimated correlation coefficient, $\hat{\rho}$, between the two types of measurements, and the estimated standard deviation of the expensive measurements.

The correlation and standard deviation are computed so that the VSP user can evaluate if the value of those parameters that were entered into the VSP DQO dialog box are valid. If not, the new values can be entered into VSP to obtain revised values of n' and n. VSP also produces a regression plot of the inexpensive and expensive measurements so the user can graphically evaluate the linear regression assumption. Also, VSP provides a

warning to the user that the computed confidence interval may be too short if n' and n are very small.

If CS is not cost effective, then VSP assumes simple random sampling and only the expensive measurement method will be used. VSP computes the required number of samples, n, using the iterative procedure in Gilbert (1987, page 30). Once the n expensive measurements are entered into VSP, then VSP computes the confidence interval assuming the data are normally distributed, i.e., by using the t distribution with n-1 degrees of freedom.

4.0 Test if the Mean Exceeds a Fixed Threshold Value

Suppose the sampling objective is to estimate the mean and conduct a one-sample test of the null hypothesis that the mean exceeds a fixed threshold value. The methodology for computing n' and n needed for the test has recently been developed by the authors and is currently being coded into VSP. The VSP user will access the dialog box for this methodology by clicking **Sampling Goals > Compare Average to Fixed Threshold > Can Assume Data will be Normally Distributed > Collaborative Sampling > Simple Random Sampling** or **Systematic Grid Sampling.**

First, the VSP user inputs the following data quality objective into the VSP dialog box: the null hypothesis of interest ("true mean ≥ threshold value," or "true mean ≤ threshold value," the tolerable probability, α, that the test will falsely reject the null hypothesis, the tolerable probability, β, that the test will falsely accept the null hypothesis, the width of the gray region, Δ, in the Decision Performance Goal Diagram, the expected total standard deviation of the set of expensive measurements, $\sigma_{total,ex}$, the expected correlation, ρ, between the inexpensive and expensive measurements, and the measurements costs C_{ex} and C_{inex}.

Then VSP uses Equation 1 above to determine if CS is cost effective relative to simple random sampling.

If CS is cost effective, then VSP computes n' and n using the following equations, which were derived by the authors using the method of proof used in Appendix A in EPA (2000b)

$$n' = \left[\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2_{total,ex}}{\Delta^2} + \frac{1}{2} z^2_{1-\alpha} \right] \rho\left(\sqrt{R\left(1-\rho^2\right)} + \rho\right)$$

$$n = \left[\frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2 \sigma^2_{total,ex}}{\Delta^2} + \frac{1}{2} z^2_{1-\alpha} \right] \left[1 - \rho^2 + \rho\sqrt{\frac{\left(1-\rho^2\right)}{R}} \right]$$

After the n' inexpensive and n expensive measurements are obtained and entered into VSP, then VSP computes

- The mean, $\bar{y}_{cs}$, and it's variance, $s^2_{\bar{y}_{cx}}$, using Equations 9.1 and 9.2, respectively, in Gilbert (1987, page 107)
- The correlation coefficient between the two types of measurements and the standard deviation of the expensive measurement, and
- A one-sample Z test of the null hypothesis

If the VSP user selected the null hypothesis to be "true mean exceeds the fixed threshold value," then the Z test is conducted by computing

$$Z = \frac{\bar{y}_{cx} - ThresholdValue}{s_{\bar{y}_{cs}}}$$

and rejecting the null hypothesis if $Z \le -z_{1-\alpha}$, where $z_{1-\alpha}$ is the (1-α)th percentile of the standard normal distribution. The Z test is used instead of the t test because the most appropriate method for determining the degrees of freedom for the t test for the CS design has not yet been determined.

VSP also constructs a regression plot of the two types of measurements, and if both n' and n are small, warns the user that the test result may not be reliable. Finally, VSP automatically reduces the value of ρ entered in the dialog box by 0.10 units (say from 0.80 specified in the dialog box down to 0.70) and re-computes n' and n. This permits the VSP user to see how n' and n change if the original value of ρ was too large by 0.10. VSP also conducts a sensitivity analysis to determine how n' and n are affected by changing the DQO input parameters. This sensitivity analysis is provided in the automatically-generated design report. This report may be inserted in a Quality Assurance Project Plan or similar project documents.

If CS is <u>not</u> cost effective, then VSP computes the number of expensive measurements, n, needed to test the null hypothesis using the following equation (derived in Appendix A of EPA 2000b), which is suitable if simple random sampling is used:

$$n = \frac{\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2 \sigma^2_{total,ex}}{\Delta^2} + \frac{1}{2} Z^2_{1-\alpha}$$

After the n expensive measurements are obtained, VSP computes the mean and its standard deviation using standard statistical formulas appropriate for simple random sampling (not CS). Finally, VSP performs a one-sample Z test and reports whether the null hypothesis can be rejected at the α-significance level.

An example of some VSP output when CS is cost efficient is shown in Figure 1.0, which shows the VSP dialog box and DQO inputs, the resulting number of samples (n' and n) computed by VSP, and the sampling locations placed on the map of the site.

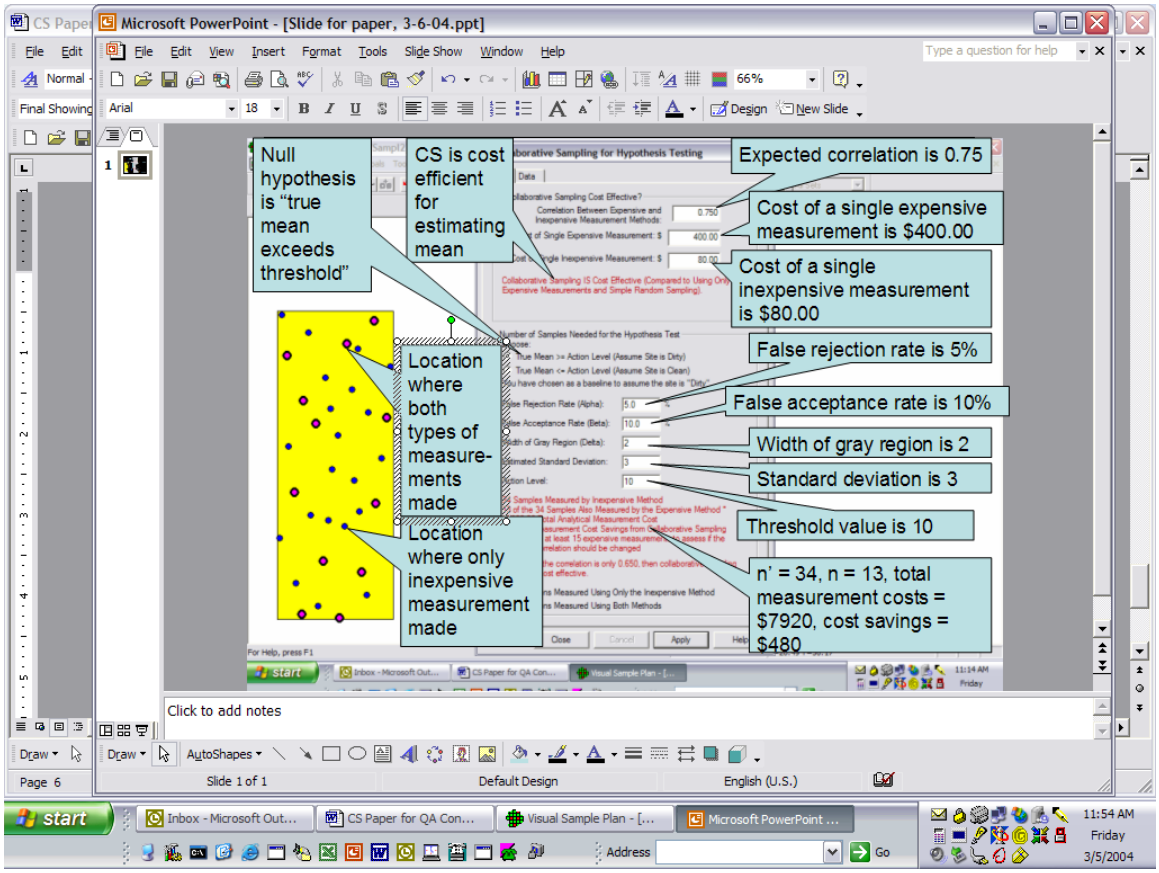


Figure 1. Example VSP Dialog Box and Map for Hypothesis Testing

5.0 References

EPA 2000a. *Guidance for Data Quality Assessment, Practical Methods for Data Analysis, EPA QA/G-9*, EPA/600/R-96/084, U.S. Environmental Protection Agency, Office of Environmental Information, Washington, DC.

EPA 2000b. *Guidance for the Data Quality Objectives Process*, EPA QA/G-4, EPA/600/R-96/055, U.S. Environmental Protection Agency, Office of Environmental Information, Washington, DC.

EPA 2002. *Guidance on Choosing a Sampling Design for Environmental Data Collection , EPA QA/G-5S,* EPA/240/R-02/005, U.S. Environmental Protection Agency, Office of Environmental Information, Washington, DC.

Gilbert, R.O. 1987. *Statistical Methods for Environmental Pollution Monitoring*, Wiley & Sons, New York.

Determining Detection Limits for Environmental Analyses

Thomas Georgian, U.S. Army Corps of Engineers

A simple cost-effective approach is proposed for estimating detection limits in lieu of the method detection limit (MDL) procedure in 40 CFR Part 136 Appendix B. Unlike the MDL procedure, the approach is protective against false positives (e.g., since it takes long-term variability into account), addresses false negatives, and takes analytical bias into account.

Introduction:

In March 2003, the U.S. Environmental Protection Agency (EPA) proposed revisions for the method detection limit (MDL) procedure in 40 CFR Part 136 Appendix B[1]. Although the MDL procedure was established for determining the sensitivity of analytical test methods under EPA's Clean Water Act (CWA), the procedure has become the de facto industry standard for determining detection capability for environmental test methods. EPA's re-assessment of the existing MDL procedure and the proposed revisions that resulted from the evaluation were performed "pursuant to a settlement agreement with the Alliance of Automobile Manufacturers, *et al."* The lawsuit challenged the validity of the "the general procedures used to establish the method detection limit." Unfortunately, as stated by, the American Council of Independent Laboratories (ACIL), "The proposed approach is fundamentally an extension of the previous application and remains a poor method to determine a laboratory's method sensitivity"[2]. Some salient problems with the existing as well as the proposed MDL procedure are discussed below:

- The MDL underestimates long-term analytical variability. Since the MDL is typically determined from a single analytical event (e.g., a set of replicates samples processed in the same batch and analyzed the same day), the MDL procedure does not take into account analytical variability that arises from different analysts, instrument calibrations, lots of reagents, and so forth. For example, in a study by the ACIL that compared the variability of "long-term method blanks for method 200.7" and "method blanks prepared on the same day analyzed in a single batch," it was observed that "The standard deviation of the long-term method blanks was typically 2 to 4 times greater than the single batch method blanks…[2]"

- The MDL is not a conservative statistical limit for the minimization of false positives. It protects against false positives (with 99% confidence), but only for the next single future measurement and does not protective against false positives for a large unspecified number of future measurements [3]. The MDL is typically determined from only seven replicate measurements, giving rise to a statistical estimate of the detection limit that can vary from one determination to the next (e.g., even over a short time period) by a factor of about two.

- The MDL does not address false negatives. The proposed definition is as follows: "The method detection (MDL) is an estimate of the measured concentration at which there is 99% confidence that a given analyte is present in a given matrix" (which is not substantively different from the current definition: "…the minimum concentration of a substance that can be measured and reported with 99% confidence that the analyte concentration is greater than zero…"). The MDL is primarily calculated using the formula [1, 4]:

$$MDL = t_{1-p,n-1}\, s \qquad (1)$$

The factor $t_{1-p,n-1}$ denotes the $(1-p)$ 100^{th} percentile of the Student t distribution with $n-1$ degrees of freedom, where p = 0.01; s denotes the standard deviation for a set of n replicate measurements (where $n \geq 7$). Therefore, the MDL is essentially a "critical value" (i.e., a limit that minimizes false positives). If an analyte were not present in a sample, the probability that a *single* future sample measurement would be less than the MDL would be about 99%. Thus, if a measurement were greater than the MDL, the analyte would be reported as "detected" with at least 99% confidence. However, if a measurement were less than the MDL, the result could not be reported as "< MDL" with a high level of confidence. If the analyte were present at a concentration near the MDL, for example, the result would be erroneously reported as "< MDL" (a false negative) about 50% of the time [3].

- The MDL does not take analytical bias (positive or negative) into account because it is calculated on the basis of only analytical precision. Neither upper nor lower acceptance limits are established for analytical bias. For example, it is primarily assumed that the mean concentration of a blank is zero (e.g., the revised MDL procedure does not permit "a recovery – or blank-correction procedures" unless it is specified in the test method). As a result, the procedure under estimates detection limits for test methods that possess positive bias at low-concentrations due to persistent blank contamination. The MDL could be substantially smaller than the mean analyte concentration in method blanks [2,5].

- The MDL procedure can be a relatively expensive to perform because of the number of replicates potentially required to determine the MDL for each analytical method and instrument. For example, the proposed procedure states: "When developing an MDL for a new or revised method, or when developing a matrix-specific MDL, the MDL procedure must be iterated and the reasonableness of the MDL determined using an F-test…" If the ratio of the variances for the two MDL determinations exceeds a specified critical value for the F-test, then yet another iteration of the MDL procedure must be performed until the critical value for the F-test is not exceeded. The approach increases the cost of the performing MDL studies by at least a factor of two.

This primary objective of this article is to present an alternative procedure for calculating detection limits that overcomes most of the shortcomings of the both the current and proposed EPA MDL

procedures (e.g., the issues discussed above), while remaining relatively cost-effective for routine environmental production work. The procedure is essentially a more complete treatment of detection limits presented in a prior publication [3]. Like the 40 CFR procedure, the approach uses a single-concentration-based design. A calibration-based design would be expected to produce superior results, but would not be as cost effective for routine environmental testing (e.g., since more analyses would typically be required to establish detection capability).

As in Currie's approach, two "types" of detection limits (DLs) are proposed: A "critical value" or detection limit that minimizes Type I error (i.e., "false positives"), denoted as L_C (using Currie's notation) and a detection limit that minimizes Type II error ("false negatives"), denoted as L_D [8]. The "Type I detection limit," L_C, is essentially a reporting limit for "detections," and the "Type II detection limit," L_D, a reporting limit for "non-detections." A measured concentration, X, would be reported as "detected" (at the 99% level of confidence) if $X > L_C$ and all method-specified identification criteria were met (peak pattern recognition occurs for the Aroclor analyses by Method 8082, and all qualifier ions are present for volatile analyses by Method 8260B). Otherwise, the result would be reported as a "non-detect." The limit L_D defines the lowest reporting limit for non-detects. Thus, if $X < L_C$ or method-specific identification criteria are not met, the result would be reported as "$< L_D$" (or less than some threshold value greater than L_D).

Procedure:

A detection limit study is performed as part of an "Initial Demonstration of Proficiency" (IDOP). If blank contamination is not significant, the IDOP is essentially performed using the 40 CFR Part 136 procedure. At least seven low-level laboratory control samples (e.g., method blanks spiked at a concentration from 2 to 10 times the estimated analytical detection limit) are processed through the entire test method (i.e., through all preparatory and determinative steps). However, the seven replicates are analyzed in at least three separate analytical batches over three or more days to help take day-to-day sources of analytical variability into account. The replicate low-level laboratory control samples (LCSs) are used to calculate the sample standard deviation, *s*. For methods capable of reporting uncensored numerical results (e.g., ICP trace metal methods), L_C may be calculated using at least seven method blanks. However, since method blanks are routinely analyzed as batch quality control samples, a large number of method blanks (e.g., at least 30) is recommended to calculate *s,* since the ultimate objective is to produce an accurate estimate of the "true" (i.e., population) standard deviation, σ. Additional LCSs (e.g., analyzed on a per batch basis) could also be used. The sample statistic *s* becomes a better estimate of σ as the number of replicates n increases.

If fewer than 20 or 30 data points are available (the typical scenario), the Type I detection limit is calculated using the formula [3]:

$$L_C = z_{1-p}\, s_{UCL,1-\gamma} = z_{1-p}\sqrt{[(n-1)/\chi^2_{n-1,\gamma}]}\; s \qquad (2)$$

where $\chi^2_{n-1,\gamma}$ is the $\gamma 100^{th}$ percentile of the χ^2 distribution and z_{1-p} denotes the $(1-p)100^{th}$ percentile of the standard normal distribution. (Note that, as in 40 CFR Part 136, normality is assumed.) Equation 2 is the "$(1-\gamma)100\%$ tolerance interval that contains at least the proportion $1-p$ of the population." If a large number of analyses were performed using method blanks, then $(1-p)100\%$ of the measurements would less than L_C with $(1-\gamma)$ 100% confidence. If $p = 0.01$ and $\gamma = 0.05$ (or 0.01), 99% percent of all future measurements will be less than L_C with 95% (or 99%) confidence. For $p = 0.01$, $\gamma = 0.01$, and $n = 7$,

$$L_C = 2.33\sqrt{(7-1)/0.872}\ s = 6.11\ s \approx 2\,MDL \qquad (2b)$$

Equation (2b) may be used in lieu of Equation (2) and provides a conservative estimate of the Type I DL when $n > 7$. (The Type I DL calculated from Equation 2 will be greater than or approximately equal to two times the MDL from Equation 1 when $n > 7$). For a large number of replicates (e.g., $n > 30$), the formula in 40 CFR Part 136 may be used to calculate the Type I DL:

$$L_C = t_{1-p,n-1}\,s \qquad (3)$$

However, it is recommended that Equation 2 be used even when $n > 30$ since Equation (3) tends to under estimate the Type I DL. It is also recommended that a statistical test for normality as well as a test for outliers (e.g., the Grubbs test) be performed prior to calculating L_C.

The Type I DL is verified using a "detection limit check sample" or "false negative quality control sample" (FNQS) rather than using the iterative F-test procedure presented in 40 CFR Part 136. *The FNQS not only verifies L_C, but also establishes L_D (i.e., becomes the lowest possible reporting limit for non-detects)*. The FNQS is prepared in the same manner as the environmental samples. For example, for drinking water, the FNQS would be reagent water fortified with the analyte(s) of concern at about 2 – 3 times the calculated Type I DL and would be processed though the entire analytical method. However, note that when many analytes are being simultaneously analyzed, it may not be practical to prepare a spiking solution that is 2 to 3 times the calculated Type I DL for all the analytes. Under these circumstances, the spiking concentration for the IDOP-DL study may be used for the Type II detection limit (FNQS) as long as it is 2 -10 times L_C.

The FNQS verifies the Type I DL. "Detection" occurs if the measured concentration, $X > L_C$ and all method-specific identification criteria are met (e.g., peak pattern recognition for the PCB analyses). Once the FNQS is analyzed, it may be necessary to analyze additional FNQS at higher or lower spiking concentrations. If an analyte in the FNQS is not detected, then L_C could have been under estimated or a low bias could be present. The concentration of the FNQS must be increased until the analyte can be consistently detected (e.g., in two consecutive FNQS). If a significant negative bias were <u>not</u> present, then the calculated value of L_C obtained from Equation 2 would be rejected and L_C would

be estimated using the FNQS spiking concentration; L_C would be set at one half the concentration of the FNQS. However, if the non-detect for the FNQS were to arise from a large negative bias, then the calculated Type I DL would be retained, but L_D would be established from the lowest FNQS concentration that produces a detect.

If the response (i.e., signal to noise ratio) of the analyte in the FNQS is very high, then L_C may be over estimated; the concentration of the FNQS may be too high. The FNQS spiking concentration may be decreased to the smallest concentration that gives rise to detection. If two consecutive FNQS produce detections, then L_C may be reduced to one half the lowest FNQS concentration that gives rise to detection. Note that reducing the FNQS to obtain a lower value of L_C or L_D would be optional. However, if analytical bias is not a significant factor and the FNQS is greater than three times the calculated Type I DL, one half the concentration of the lowest FNQS (that gives rise to a detection) should be reported as an upper bound for L_C.

It should also be noted that when L_C calculated from Equation 2 is high relative to the value of L_D determined from the FNQS concentration (e.g., L_C could be greater than L_D), setting $L_C = L_D / 2$ should be done with caution. For example, if the calculated value of L_C (Equation 2) is based upon a large number of values and L_D is based upon one or two FNQS analyses performed within a short period of time, the FNQS analyses may not be a valid measure of L_D. Under these circumstances, FNQS data should be collected over a period of time before establishing L_D (e.g., the period of time and analytical conditions should be comparable to that for the data collected to calculate L_C).

If the FNQS verifies L_C, then the reporting limit for non-detects (i.e., L_D) should be no less than the FNQS; that is, non-detects should be reported as "< Y," were "Y" denotes the concentration of the FNQS. However, it should be noted that when statistical evaluations of the data will be performed, any degree of data censoring is often undesirable. It is recommended that all numerical results be reported with the values of L_C and L_D (e.g., a non-detect could be reported as "< Y [X]," where X denotes the measured value). One FNQS should be analyzed periodically (e.g., one per analytical batch, initial calibration, or weekly, depending on the nature of the method) as an on-going demonstration of L_C. Minimally, quarterly analyses are recommended.

Optionally, as an additional check, analyze a FNQS per batch and calculate the recovery of the FNQS. After 20 FNQS results have been collected, calculate upper and lower control limits for the recoveries. Until 20 FNQS results have been collected, initial control limits of +/-50% are recommended. (The RSD at two times the DL from Equation 3 is about 20% when $n = 20$ since t = 2.53 and 100/(2 x 2.53) = 20%; if limits are set at 2-sigma, then 2 x 20% = 40%.) Note any FNQS results outside of control limits. If an excessive number of FNQS results are out side of control limits (e.g., more than 10%), the Type I DL may have increased since the initial determination and should be recalculated. L_C has significantly increased (and should be revised) if the new value of L_C is greater than two times the original.

Statistical tests that compare the degree of dispersion (precision) between two or more data sets could be used to monitor significant changes in L_C. For example, Levene's test could be used to compare the variance of a data set used to calculate an initial value for L_C to the variance of a data set used to calculate a revised value for L_C. Levene's test is similar to Bartlett's test (F-test) for the homogeneity of variances, but is more robust to departures from normality (e.g., since it evaluates dispersion about the median rather than the mean) [6]. Like the Bartlett's test, if the test statistic exceeds a specified critical value, the "baseline" assumption (null hypothesis) that the variances of the two data sets are equal is rejected. If Levene's test were to indicate the variances are significantly different (e.g., the 95% or 99% level of confidence), one would conclude that a significant change in L_C has occurred. A new L_C value would be calculated (e.g., using Equation 2) and verified using a FNQS as discussed above. *Note that, regardless of whether a statistical test is used to monitor changes in L_C, the use of a FNQS to verify L_C is essential* (e.g., since this takes changes in analytical sensitivity due to bias into account; good precision but extreme negative bias would result in a failure to detect the analyte in the FNQS).

For analytical methods characterized by low-level blank contamination that cannot be completely eliminated, L_C is defined as the concentration that is statistically different from a method blank at the 99% level of confidence. Method blanks are used to calculate L_C when the mean blank concentration is significantly different from zero. At least seven method blanks (*not* spiked with any analytes) are analyzed. For *n* replicates, the Type I DL would be the 99% or 95% upper tolerance limit ($\gamma = 0.99$ or 0.95) for 99% coverage ($p = 0.99$):

$$L_C = \bar{x} + K_{\gamma,p,n}\, s \qquad (4)$$

The quantity $\bar{x}$ denotes the mean concentration for the method blank. (Note, once again, normality is being assumed). The value of *K* can be obtained from tables (for the noncentral t distribution). Note that if the population mean is known and equal to zero, the above equation reduces to Equation (2). The value *K* in Equation 4 can also be estimated using the formula [7]:

$$K_{\gamma,p,n} \approx \frac{z_{1-p} + \sqrt{z^2_{1-p} - a\,b}}{a}, \text{ where } a = 1 - \frac{z^2_{1-\gamma}}{2(n-1)} \text{ and } b = z^2_{1-p} - \frac{z^2_{1-\gamma}}{n} \qquad (5)$$

Note that when calculating L_C (especially using blanks), it is acceptable for some results to be negative values. A simple conservative approach for establishing L_D would consist of multiplying the value of L_C calculated using Equation 4 by a factor of two. Alternatively, L_D could be established via the analysis of a FNQS as discussed above (spikes at L_D would be required to give rise to a measured value greater than L_C).

Discussion:

The proposed approach possesses a number of advantages over the 40 CFR Part 136 MDL procedure. The approach gives rise to a Type I DL that is more protective of false positives than the 40 CFR Part 136 MDL procedure since it more effectively takes long-term analytical variability into account and is based upon the tolerance interval for an unspecified number of future observations rather than prediction interval for the next single future observation. Furthermore, unlike the MDL, the approach addresses false negatives. The FNQS establishes the Type II DL, the lowest reporting limit for non-detections, and provides an empirical verification of the Type I DL. Unlike the MDL procedure, the Type I DL is determined during method development only and is verified periodically via the analysis of a FNQS. Control charts can also be optionally generated for the FNQS recoveries to more effectively monitor method performance (e.g., to identify possible changes in method sensitivity).

References:

1. U.S. Environmental Protection Agency. *Federal Register*, *68, No. 48*, March 2003, 11770 - 11793.
2. American Council of Independent Laboratories. *Environmental Laboratory Washington Report*, Vol. 14, Issue 8, April 2003.
3. Thomas Georgian and Kenneth Osborn. "Determining 'Reliable Detection Limits' (RDLs) for Environmental Analyses, *Quality Assura*nce, 8:1-9, 2003.
4. U.S. Environmental Protection Agency. *Federal Register*, 49, 1984, 43430-43431.
5. American Council of Independent Laboratories. "An Evaluation of the EPA's Revised Procedure for Method Detection Limits," ACIL Environmental Services Section Technical Committee memorandum, Revision 1.0, March 20003.
6. Brown, M. B. and Forsythe, A. B. "Robust Tests for the Equality of Variances." *Journal of the American Statistical Association,* 69, 1974, pp. 364–367.
7. Owen, D. B. "A survey of Properties and Applications of the Noncentral t-Distribution," *Technometrics*, Vol. 10, No. 3, August 1968.
8. Currie, A. L. "Nomenclature in Evaluation of Analytical Methods Including Detection and Quantification Capabilities, *Pure and Applied Chemistry*, Vol. 67, No. 10, pp. 1699-1723, 1995.

A Statistical Methodology for Estimating Background Concentrations

Basil Coutant, Battelle

A common problem in environmental monitoring is establishing the mean background concentration of a pollutant at a site that is affected by local sources. These background estimates may be subtracted off to establish the local contribution or they may be the basis for target concentrations of future pollutant levels. Traditionally the background concentrations have been estimated by isolating sampling periods when it is felt that local sources do not affect the site or by pairing the site with another site that is assumed to measure only the background concentration for the primary site of interest. There are problems with both of these approaches. This paper presents a statistical methodology for estimating background concentration that utilizes all (or most of) the data collected at the site. The modeling is based on maximum likelihood estimation with a distribution derived by modifying a gamma distribution. The modifications to the Gamma distribution address detection limit issues and can be used to estimate background levels near or below the detection limit. The method is demonstrated with benzene data collected in the Portland, Oregon, area and is comparable to more traditional estimates of background concentration.

Introduction

The goal for this project was to estimate annual (or typical) background concentrations for ambient concentration measurements. By definition, the approach should not seek to identify what occurs during exceptional events. Rather, the desired approach needs to identify the typical background for a site. With this perspective in mind, approaches that rely on severely restricting the monitoring data to a small subset of observations corresponding to certain events (e.g., days with persistent winds, sharp frontal passages, or the right wind trajectory) should be rejected. Instead it should be acknowledged that evidence of background concentration levels is likely contained within all measurements, and this information should be exploited to estimate the background.

To do so this paper[1] represents monitoring data by a statistical model developed in several stages. The basis for the statistical model used is a gamma distribution[2], described mathematically by its probability density function (pdf).

In general, an ordinary gamma distribution has support on the real interval $[0, \infty)$. That is, it applies to variables with a non-negative range of values. The gamma distribution is defined by two parameters, a shape parameter α and a scale parameter β. The specific values of the parameters α and β impact the appearance of the gamma pdf, and variations

[1] Battelle's work was partially supported under EPA Contract No. 68-D-02-061.

[2] The log-normal distribution was also considered, but did not fit the test cases as well. This may be due to a numerical stability issue for a shifted log-normal distribution that is not shared with the shifted gamma distribution.

of these parameters provide for a very flexible family of data-modeling distributions. A shifted gamma distribution introduces a third parameter, call it μ, which changes the support of the distribution from that of the ordinary gamma, i.e., from $[0, \infty)$ to $[\mu, \infty)$. The basic shape of the gamma pdf is left unchanged by such a shift. In the current context, μ is assumed to be non-negative; however, this constraint is not required in general.

To illustrate the model development and support its logic, consider the benzene concentrations observed at five monitoring sites in Portland, Oregon, from July 1999 through July 2000. Figure 1 summarizes the data via overlaid, site-specific time series plots. While there are high and low concentration periods, there is no obvious seasonal trend in the Portland data. Notice that there is a distinct lack of data below 0.3 $\mu g/m^3$, about half of the data are between 0.3 and 2 $\mu g/m^3$, and the remainder of the data are spread out progressively thinner from 2 to 9 $\mu g/m^3$. Upon first glance, this behavior would appear consistent with a conceptual viewpoint and statistical approach of a constant background and a distribution of source-oriented additions to that background.

Figure 2 presents a quantile-quantile (Q-Q) plot of a gamma distribution fit to the benzene data at the Downtown site. In general, Q-Q plots demonstrate the fit, or lack thereof, of a proposed statistical distribution to the empirical behavior of a given set of data. Straight lines in such plots are indicative of a good fit. Figure 2 supports the assertion that the shifted gamma pdf is, in fact, a reasonable choice for the statistical distribution. It also shows that the data (all the data) contain a positive shift, or background. All of the data for this site are shifted up approximately 0.75 $\mu g/m^3$ from a line through the origin. This positive shift is the background in the proposed model. (See Table 1 also.)

The parameters were fit to the distributional model using the method of maximum likelihood estimation. A detailed description of maximum likelihood estimation may be found in Ref. [1]. The method is based on the probabilistic structure of the model. The parameters are fit (estimated) through an iterative procedure using the NLMIXED procedure in the SAS® software system that optimizes an object function. The object function, called a likelihood, is a mathematical description of the probabilistic structure of the data.

The likelihood has the same formula as the data's assumed probability density, but with a different interpretation. As a pdf, the parameters that are to be estimated are treated as fixed constants and the data treated as random variables. In that setting, the formula describes the probability of observing data in any given range. As a likelihood, the data are treated as fixed constants (i.e., the actual data observed via monitoring are fixed and known once observed) and the parameters are treated as variables. The maximum likelihood estimates are the ones that maximize the likelihood, and essentially represent the parameter values that would assign the highest probability to the observed monitoring outcome.

Notice in Figure 2 that the "straight line" behavior of the Q-Q curve breaks down slightly near the intercept. Since the goal is to capture the mean long-term vertical shift rather than the minimum shift, the model was modified from a shifted gamma, to a model that treats values near and below the mean shift different from the remainder of the data. The data near [within 2 times the minimum detection limit (MDL)] or below the shift are treated as random noise. The adjustment introduced to make this modification has several advantages. First, the model will naturally handle below MDL data without additional modifications. Second, the model is continuous and always supported on $[0, \infty)$. This results in a model that satisfies regularity conditions [1] so that the standard errors can be estimated via large sample theory.

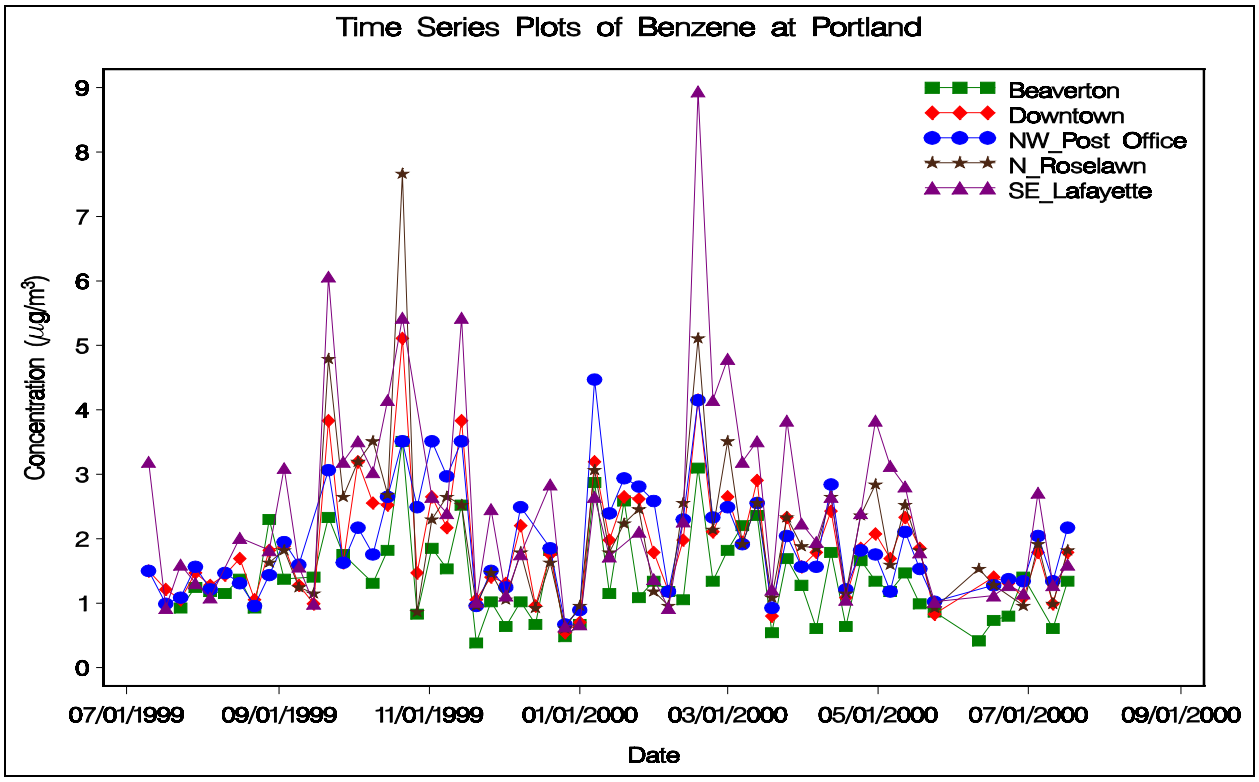


Figure 1. Benzene Monitoring (µg/m^3) from Five Portland, Oregon, Monitoring Stations Operating from July 1999 through July 2000.

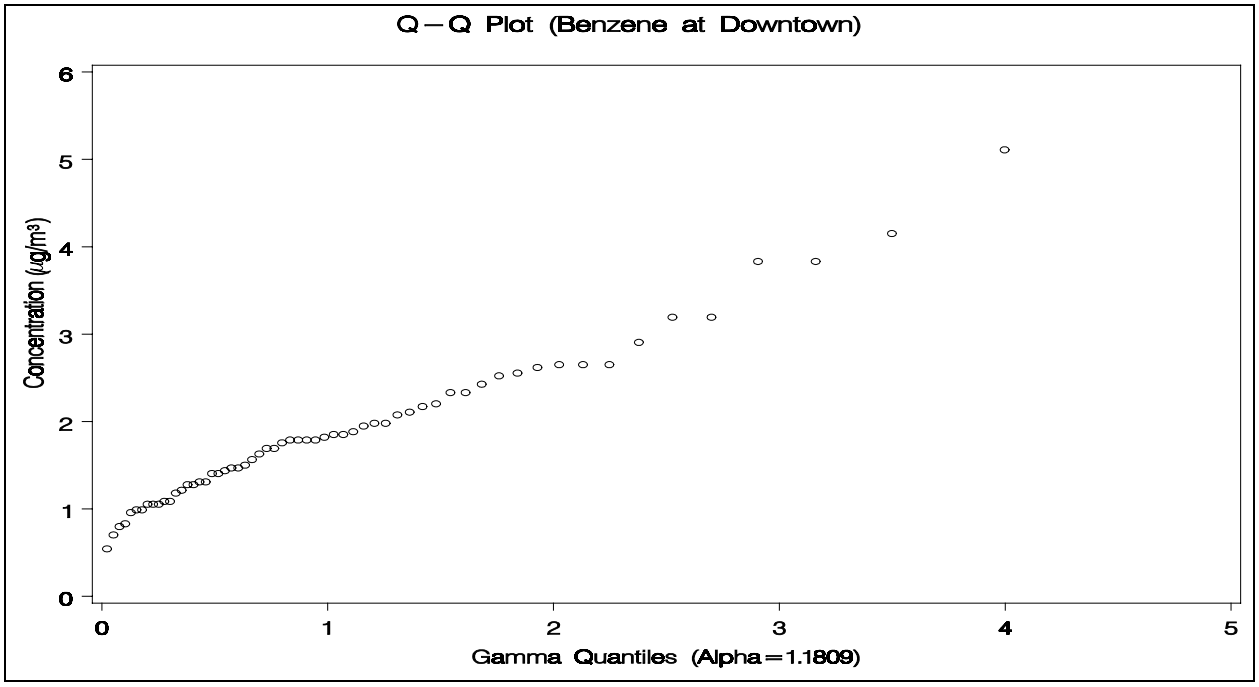


Figure 2. A Quantile-Quantile Plot of a Gamma Distribution Fit to Benzene Data from the Downtown Site in Portland, Oregon (July 1999-July 2000).

The development of the model starts with the shifted gamma distribution. The form of such a model and likelihood/distribution may be written as follows:

$$L(\mu, \alpha, \beta) = \text{“shifted gamma”} \text{ (i.e., } y_i = \mu + \varepsilon_i \text{)} , \qquad \text{(Model 1)}$$

where y_i represents the i^{th} individual concentration; μ represents the true unknown background concentration level; and ε_i represents short-term, source-oriented “shocks” that produce positive deviations from the long-term background. The ε_i’s are assumed to behave according to an ordinary gamma distribution with parameters α and β, implying the y_i’s have a shifted gamma distribution.

The first modification to the above model is motivated by measurement issues and numerical issues with fitting the data. Frequently, MDLs are defined in terms of a measurement. For modeling purposes here, such a point of view is adopted. Data within a threshold of background plus two times the reported MDL (μ+2*MDL) were treated as random noise, or at least too imprecise to use individually for estimating the parameters of the assumed shifted gamma distribution. The MDL for the Portland data shown in Figure 1 is 0.1 $\mu g/m^3$, so the target for the threshold is about 0.95$\mu g/m^3$. In effect, all such data are censored and treated simply as an indicator of “below a threshold.” Although non-numerical in nature, such censored data are still used in the statistical

modeling process by making the proportion of the data below the threshold consistent with the data above the threshold.

Now, the data are treated in a dichotomous fashion depending on the relationship to the censoring threshold. Specifically, Model (1) is first modified as follows:

$$L(\mu,\alpha,\beta)=\begin{cases}\text{"shifted gamma"} & \text{if } y_i \geq \mu + 2\cdot MDL \\ \text{"constant"} & \text{otherwise}\end{cases} \qquad \text{(Model 2)}$$

where y_i and μ are defined to be the same as in Model (1).

The decision to censor data up to a threshold of (μ+2*MDL) eliminated the problem of needing to know if the censoring threshold was larger or smaller than the background before fitting the models. By using (μ+2*MDL), the censoring threshold is always greater than the background.

While the above model can be fit numerically, the dichotomous treatment of the data results in a discontinuity in the likelihood. This discontinuity at (μ+2*MDL) can cause a numerical instability in the estimate of the standard error of the background parameter. Hence, a further modification to the likelihood was introduced. The modification introduced was to connect the two portions of the likelihood with a positive, finitely-sloped line over a short range; namely, a range of 1 MDL. This ensures that the object function used in the parameter estimation is continuous and generally ensures that the uncertainty is estimable[3].

$$L(\mu,\alpha,\beta)=\begin{cases}\text{"shifted gamma"} & \text{if } y_i \geq \mu + 2\cdot MDL \\ \text{"linear"} & \text{if } \mu + MDL \leq y_i < \mu + 2\cdot MDL \\ \text{"constant"} & \text{otherwise}\end{cases} \qquad \text{(Model 3)}$$

Mathematical Details

There are two mathematical details that need to be explored. First, what does forcing the likelihood to be continuous really gain, and second, how is the constant found?

Forcing the likelihood to be continuous causes the likelihood to satisfy a regularity condition, namely that one can differentiate under the integral. To see why this works, consider a hypothetical case where *f* is a continuous pdf with the form:

[3] For the sake of theory, one should replace Model 3 with a model that is exactly the same except that the two ends of the linear part are replaced with cubic splines so that the likelihood has a continuous first derivative on (0, ∞). In theory this can be done very easily and in such a way that the new model is the same as Model 3 to within the precision of a computer. This final model yields exactly the same estimates as Model 3 and satisfies the full regularity conditions. Hence, the maximum likelihood standard errors given by SAS for the Model 3 parameter estimates are consistent for large samples. Clearly, there is no point in working with the more complex model in practice.

$$f(x,\gamma)=\begin{cases} g(x,\gamma) & if\ x \le r(\gamma) \\ h(x,\gamma) & if\ r(\gamma) < x \end{cases}.$$

Then for a continuous s(x) with finite expectation:

$$\frac{d}{d\gamma}\int_{-\infty}^{\infty} s(x)\cdot f(x,\gamma)dx = \frac{d}{d\gamma}\int_{-\infty}^{r(\gamma)} s(x)\cdot f(x,\gamma)dx + \frac{d}{d\gamma}\int_{r(\gamma)}^{\infty} s(x)\cdot f(x,\gamma)dx$$
$$= \int_{-\infty}^{r(\gamma)} s(x)\cdot\frac{\partial}{\partial\gamma} f(x,\gamma)dx + s(r(\gamma))g(r(\gamma),\gamma) + \int_{r(\gamma)}^{\infty} s(x)\cdot\frac{\partial}{\partial\gamma} f(x,\gamma)dx - s(r(\gamma))h(r(\gamma),\gamma)$$
$$= \int_{-\infty}^{\infty} s(x)\cdot\frac{\partial}{\partial\gamma} f(x,\gamma)dx$$

since the continuity of *f* means that g(*r*(γ),γ) = h(*r*(γ),γ). Model 3 takes advantage of this by "jumping" from the constant to the shifted gamma over a range with a fixed length of 1 MDL.

The final step is to derive the mathematical form of the likelihood function. The form for the shifted gamma is the same as an ordinary gamma evaluated at (x - μ). The other two parts must be derived. The conditions required are (1) the total area under the curve must be one and (2) the function should be continuous at the two break points. Using the continuity conditions, the first condition becomes:

$$1 = \int_{0}^{\mu+MDL} kdt + \int_{\mu+MDL}^{\mu+2\cdot MDL} line\ dt + \int_{\mu+2\cdot MDL}^{\infty} Shifted\ Gamma\ pdf\ dt \quad \text{or}$$
$$1 = k\cdot(\mu+MDL) + \frac{1}{2}(k+y)\cdot MDL + (1-CDF(2\cdot MDL))$$

where k is the value of the constant, y is the value of the shifted gamma pdf at (μ+2*MDL) or the standard gamma pdf at 2MDL, and the CDF(2MDL) is the value of the standard Gamma CDF at 2MDL. Solving for k yields:

$$k = \frac{2\cdot CDF(2\cdot MDL) - y\cdot MDL}{(2\cdot\mu + 3\cdot MDL)}.$$

The CDF of the gamma function is available within SAS (and NLMIXED in particular). So the whole likelihood can be written in SAS. Note that for α > 1, the gamma pdf is concave down near 0. This condition ensures that k > 0.

Results

The results for fitting the model to the Portland Benzene data are shown in Table 1. The results are in excellent agreement with the intuition from the data and are comparable to the 1996 NATA background estimate of 0.48 $\mu g/m^3$. Note that the Beaverton site is located in an adjacent, more rural county and is separated from the other sites by a low ridge of hills.

Table 1. Numerical Summary of July 1999 through July 2000 Portland, Oregon, Benzene Monitoring Data ($\mu g/m^3$) and Background Modeling Results

Site	Sample Size	Mean	Standard Deviation	Max	Min	Background Estimate	Standard Error
Beaverton	56	1.3840	0.6948	3.5127	0.3832	0.4067	0.2034
Downtown	60	1.8953	0.8908	5.1094	0.5429	0.5491	0.0593
NW_Post Office	59	1.9204	0.8738	4.4707	0.1000	0.7359	0.0563
N_Roselawn	52	2.0972	1.2321	7.6641	0.6067	0.7127	0.0644
SE_Lafayette	55	2.4844	1.5601	8.9415	0.6387	0.7364	0.0784
All Sites	282	1.9511	1.1342	8.9415	0.1000	0.6282	0.0483

Reference

Cassella, G., and Berger, R. L. (1990). Statistical Inference, Duxbury Press, Belmont, California.

Overdispersion Models for the Violation of Nitrate Concentration Limits

Nagaraj K. Neerchal, Deparment of Mathematic and Statistics, UMBC
Earl A. Greene, United States Geological Survey
Minglei Liu, Department of Mathematics and Statistics, UMBC

Abstract:
The objective of this paper is to investigate possible approaches for the spatial analysis of the number of violations when the nitrate concentration exceeded beyond the regulatory limits. The study will follow up on an ordinary logistic regression analysis where the probability that the nitrate level exceeds a given level is related to the explanatory variables such as land use pattern within a radius of the well and the its geology type. We consider a spatial analysis where the basic unit of analysis is a watershed. We study distributional properties of the number of `hot' [defined as a well that exceeds the regulatory nitrate limit] wells in a watershed. It is noted that the usual binomial model is not an appropriate model here because of the spatial correlation. An appropriate analysis of this data using PROC GENMOD of SAS will be illustrated. Estimation of the logistic regression analysis under the Beta-Binomial model and a finite mixture of binomials model, which give a full likelihood approach to model such data, will also be discussed.

Extended Abstract

In a previous work on the analysis of nitrate concentration data, Kelley(2001) considered data from individual wells at a number of threshold levels. In this report, we analyze data combined at the watershed levels. To facilitate presentation of the method, we will only consider the threshold 3mg/l. The analysis will apply for any threshold level. We consider the following logit model for the probability of a well exceeding 3mg/l:

$$P(T_{ji} = 1) = \frac{\exp(x_j \beta)}{1 + \exp(x_j \beta)}, \qquad (1)$$

where T_{ji}, is a 0/1 variable taking value 1 if the jth well within the ith watershed exceeds the threshold 3mg/l in nitrate concentration and the rx1 vector x_i consists of predictor variables and β denotes the vector of corresponding regression coefficients. If we assume that the wells-level data are independent, then $T_j = \sum_{i=1}^{m_i} T_{ji}$ is distributed as a binomial random variables. The maximum likelihood estimates of such a model may be obtained using PROC LOGISTIC. Results from this run are given in the attached SAS printout. In the logistic model, Atmospheric deposition (HUC_ATDEP), Depth of bedrock (ROCDEP), Water table depth (WTDEP) come out to be significant.

The key assumption of independence of wells-level data is at best a working assumption. Nitrate concentrations in the wells located in the same watershed are expected to be correlated. Therefore, the number of hot wells (denoted by, say, T) in a watershed cannot be modeled as a binomial random variable. When the wells are positively correlated, the variability in T is expected to be higher than a binomial random variable. This phenomenon is known as `overdispersion' (or Extra Variation) and has been an active research area for a number of years. A number of models are available to model such data. These include Beta-binomial model and a finite mixture of binomial distributions. See Morel and Neerchal(2000), for a review. Neerchal and Morel (1998) propose a goodness of fit test for testing the adequacy of overdispersion models for real data. The test may be used to choose an appropriate model from several choices.

Generalized Estimating Equations (GEE) is a method of estimation, which relies on the structure moments, rather than the exact specification of the distribution of the data. Overdispersed binomial counts have a particular moment structure. Therefore one may use the GEE methodology to estimate the parameters of an overdispersion model. GEE estimation can be carried out is a SAS procedure called PROC GENMOD. One would then obtain appropriate estimates and estimated standard errors, which respect the spatial correlations in the data. The GEE method of estimating a logistic regression model for the nitrate concentration data is outlined below. We illustrate that certain dependence structure can be incorporated into the logistic model given by (1) and an improved analysis can be obtained using PROC GENMOD and Generalized Estimating Equations (GEE). Under the GEE, in addition to the mean structure given by (1), we assume that the variance covariance matrix of the vector of observations $T_j = (T_{j1},.....,T_{jm_j})'$, is given by

$$Var(T_j) = V_j \qquad (2).$$

Then the GEE estimates are obtained by solving,

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{j=1}^{n} \mathbf{H}_j(\boldsymbol{\beta}) \left(\mathbf{V}_j\right)^{-1} \{\mathbf{y}_j - \boldsymbol{\mu}_j(\boldsymbol{\beta})\} = \mathbf{0} \qquad (3)$$

where $\mathbf{H}_j = \left(\mathbf{H}(\boldsymbol{\beta},\mathbf{x}_{j1}), \mathbf{H}(\boldsymbol{\beta},\mathbf{x}_{j2}),...,\mathbf{H}(\boldsymbol{\beta},\mathbf{x}_{jm_j})\right)$, $\mathbf{H}(\boldsymbol{\beta},\mathbf{x}_{ji}) = \dfrac{\partial f(\boldsymbol{\beta},\mathbf{x}_{ji})}{\partial \boldsymbol{\beta}}$ and $f(\boldsymbol{\beta},\mathbf{x}_{ji})$

is the logistic mean function given by (1). The asymptotic variance-covariance matrix of the GEE estimator is given by

$$\widetilde{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \left\{\mathbf{I}_0(\hat{\boldsymbol{\beta}})\right\}^{-1} \mathbf{I}_1(\hat{\boldsymbol{\beta}}) \left\{\mathbf{I}_0(\hat{\boldsymbol{\beta}})\right\}^{-1}, \qquad (4)$$

where

$$\mathbf{I}_0(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{n} \mathbf{H}_j\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{V}_j\right)^{-1}\mathbf{H}_j^t\left(\hat{\boldsymbol{\beta}}\right),$$

$$\mathbf{I}_1(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{n} \mathbf{d}_j\mathbf{d}_j^t \text{ , and } \mathbf{d}_j = \mathbf{H}_j\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{V}_j\right)^{-1}\left\{\mathbf{y}_j - f(x_j,\hat{\boldsymbol{\beta}})\right\}.$$

This estimator of variance-covariance matrix of the parameter estimate is known in the literature as the ``sandwich estimator''. It's well known property that it gives the correct estimates the variance-covariance matrix consistently even when the hypothesized form of V_j is incorrect, has resulted in its being called the `robust' estimate. PROC GENMOD provides several options for the within-watershed covariance matrix $V_j$. Assumption of independent well-data corresponds to taking $V_j$ to be a diagonal matrix. The option in GENMOD called `EXCHANGEABLE' assumes that the off-diagonal elements of V_j are all equal.

Full likelihood based approaches such as Beta-binomial and the finite mixture of binomials mentioned earlier are not available as options in GENMOD. However, the classical method of computing maximum likelihood estimates, namely the Fisher scoring algorithm, can be implemented as shown by Morel and Neerchal (1993). Goodness-of-fit statistics for determining if a specific model is appropriate for the given data set are discussed in Neerchal and Morel (1998). If a model were applicable and interpretable, it would be beneficial to use the model for estimation, because of the gain in efficiency.

Acknowledgments

Support for completing this report was provided by USGS, via the purchase order number 01ERSA0478. Opinions expressed are the authors alone and not the funding agency's.

References

1. Kelley, Kerri-Ann, (2001). ``Spatial Probability Modeling of Ground Water Nitrate Contamination in the Mid-Atlantic Region", Unpublished MS thesis, Department of mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD 21250.

2. Morel, J. G. and Neerchal, N. K. (1993). ``A Finite Mixture Distribution for modeling Extra Variation'', Biometrika, 80, 363-371.

3. Morel, J. G. and Neerchal, N. K. (1997). ``Cluster Binary Logistic Regression Using a Finite Mixture Distribution with Application to Teratology Experiment'', Statistics in Medicine, 2843-2853.

4. Neerchal, N. K. and Morel, J. G. (1998). ``Large Cluster Efficiency of Two Parametric Multinomial Extra Variation Models'', Journal of American Statistical Association, 1078-1087.

5. Morel, J. G. and Neerchal, N. K. (2000). Extra Variation Models'', Encyclopedia of Biopharmaceutical Statistics, 197-209.

Measuring Pesticides, Lead, Allergens, and other Dangers in Homes

John Rogers, David A. Marker, and Pam Broene, Westat
Peter J. Ashley, Warren Friedman, John H. Miller II, HUD

The U.S. Department of Housing and Urban Development (HUD) is conducting the American Healthy Homes Survey (AHHS) in 2004-5, with additional support from other agencies, including EPA. This national study will collect information on allergens, chromated copper arsenate (CCA) treated wood, carbon monoxide, formaldehyde, lead, mold, nicotine, perfluorooctanoic acid (PFOA), pesticides and child safety hazards in homes. Samples will be collected from paint, dust (by wiping and by vacuuming), soil, and air; supplemental information will be obtained through questionnaires and observations by trained interviewers/technicians. This presentation will provide details on the statistical design and sampling protocols for this study. Some of these several materials being analyzed (analytes) were also measured in the National Survey of Lead and Allergens in Homes (sponsored by HUD and NIEHS) in 1999-2000 and the First National Environmental Health Survey of Child Care Centers (HUD, EPA and CPSC) in 2001. AHHS will provide an opportunity to see how exposure to these analytes has changed over time. The data on several of these analytes will provide the first national and regional estimates of their prevalences in homes across the United States. We will also collect demographic and behavioral information that can be used to estimate exposures for children, the poor, and other populations of interest.

Overview: The U.S. Department of Housing and Urban Development's Office of Healthy Homes and Lead Hazard Control (OHHLHC/HUD) and the U.S. Environmental Protection Agency's Office of Research and Development (EPA/ORD), conduct research designed to identify, characterize, and reduce human exposures and risks to key hazardous environmental contaminants commonly found in and around the nation's residences.

HUD, working with other agencies, has conducted national surveys generating data to characterize potential exposures, risks, and other key hazards for the U.S. population. In 1989-1990, HUD, working with EPA's Office of Pollution Prevention and Toxics (OPPT), sponsored a national survey of lead-based paint (LBP) in housing. The primary objective of that survey was to estimate the prevalence of LBP in housing. In 1997-2000, HUD and National Institute of Environmental Health Sciences (NIEHS) conducted the National Survey of Lead and Allergens in Housing (NLSAH) and produced estimates for the levels and patterns of lead hazards in United States. In 2001-2002, HUD, EPA, and the Consumer Product Safety Commission (CPSC) surveyed randomly selected child care centers across the United States to characterize children's potential exposures to lead, allergens, and pesticides.

HUD and EPA (both ORD and OPPT) are now planning for a new national survey of lead hazards, allergens, pesticides, and other important environmental and safety hazards found in and around the nation's residential housing. This new survey, the *American Healthy Homes Survey* (AHHS), will assess potential residential exposures for the general population, to these key hazards. The data collected in this study will be used to develop new distributions of exposure

and risk, and to examine changes in the occurrence and magnitude of these exposures and risks over time, where baseline data is available. For the AHHS, as for the previous surveys, HUD is obtaining Office of Management and Budget approval of the information collection request, including the survey methodology and statistical design. The survey procedures will also be approved by the cognizant Institutional Review Board.

This paper describes the planned statistical design and sampling protocols for this study.

Population of Interest: The estimates described above will be generated for the roughly 106 million non-seasonal occupied housing units in the United States, with the following exceptions:

- Housing where children are not permitted to live.
- Group housing, both institutional and non-institutional.
- Vacant housing.
- Short-term housing.
- Hotels and motels.

The following subsets of the homes have sometimes been excluded from prior studies. However, they will be included in the AHHS:

- Housing built after 1977;
- Housing units in multi-family buildings; and
- Manufactured housing units, i.e., mobile homes and trailers.

Survey Design: HUD presently plans to sample 1600 housing units selected in three sampling stages, as described below.

First-Stage: Sampling PSUs: Every area in the 50 states and District of Columbia will be assigned to one of approximately 2000 primary sampling units (PSUs, or groups of counties). A stratified sample of 75 to 100 PSUs will be selected with probability proportional to PSU population.

Second Stage: Sampling Segments: Within each selected PSU, a frame of segments will be created from Census block files, where a segment consists of one or more geographically close blocks with less than 300 housing units. A sample of approximately 8 segments will be randomly selected in each PSU.

Third-Stage: Listing and Selecting Housing Units: In each sampled segment all the eligible housing units will be listed to create a sampling frame. At the third stage of sampling, an equal probability sample of four housing units will be drawn from each list using systematic sampling.

On the assumption that 50% of the selected housing units are both eligible for the survey and agree to participate, the number of responding household is expected to be about 1,600.

Recruitment: Each household in the sample will be sent an initial contact letter introducing the study and explaining the importance and advantages of participation (including an incentive payment). Interviewers will then visit each home to contact the residents, determine eligibility

and request participation. For those homes that agree to participate in the study, the interviewer will schedule an appointment for data collection, by the same interviewer joined by a technician. **Field Data Collection**: The Interviewer will be responsible for introducing the team to the occupants, obtaining informed consent, explaining the tasks involved, conducting the room inventory, and administering the Resident Questionnaire. The Technician will be responsible for all X-ray fluorescence (XRF) analyzer lead-in-paint measurements and collection of lead dust wipe and lead soil samples. The technician will be a certified lead risk assessor in accordance with EPA regulations (40 CFR 745). All other data collection activities will be split between the two field team members in a manner that reduces the time spent in the home. The team is expected to spend no more than three and a half hours at each home, on average.

Room Selection: After listing all rooms in the home, the interviewer will randomly select one room from each of the following four room strata:

- Kitchen.
- Common living area (living room, den, or family room)
- Bedroom - If one or more children age 17 and younger reside in the home, one bedroom will be randomly selected from among the bedrooms in which the children sleep. If no such children reside in the home, one bedroom will be selected randomly from all the regularly occupied bedrooms (i.e., not a guest bedroom) in the home.
- Other room – Rooms not in the three strata defined above.

The environmental measurements and samples will be collected in the selected rooms and from the building exterior. The interviewer will conduct a walkthrough survey involving questions and observations of safety features related to falls and burns in the home, including posted emergency numbers, fire extinguishers, smoke alarms, antiskid features of floor coverings, safety gates, and grab bars in bathrooms. Several tests are done during this walkthrough, including testing of smoke alarms and measurement of hot water temperature of a tub. During the walkthrough the interviewer will also collect the respondent's vacuum cleaner bag.

Lead Paint Testing: Paint will be evaluated in a non-destructive manner by XRF to determine if lead-based paint is present in the rooms sampled. The technician will test specific components in each of the four randomly selected rooms, and the exterior of the housing unit.

Dust Sample Collection for Lead: Because applicable standards exist only for lead dust sampling by the wipe method and because wipe samples correlate well with blood leads (HUD, 1995a), dust samples will be collected on window sills and floors with and without carpeting using dust wipes in accordance with HUD guidelines. One square foot templates will be used for floor samples. The entire interior sill area will be wiped for window sill samples. The wipe samples will be sent to an EPA-recognized lab for lead analysis.

Soil Collection for Lead: Within each location, samples will be collected from bare soil, i.e., not covered with grass, concrete, asphalt, or other permanent covering, if possible. If no soil is bare, soil samples will be collected from covered surfaces, if possible. Thus, soil samples may be collected from soil covered by grass or mulch, but not concrete or asphalt. Soil sampling will be conducted in accordance with core sampling procedures described in the HUD's *Guidelines for the Evaluation and Control of Lead-Based Paint Hazards in Housing*. Only the top one-half

inch of each soil core, i.e., that portion most accessible, will be included in the sample. The wipe samples will be sent to an EPA-recognized lab for lead analysis.

Dust Sample Collection for Allergens: The dust sampling method proposed for allergens is a vacuum (fitted with a fine nylon mesh tubular fabric sampler inserted in the vacuum hose wand). HUD plans to analyze the dust samples for endotoxin and dust mite, cockroach, cat, dog, and mouse allergens using enzyme-linked immunosorbent (ELISA) assays.

Dust Sample Collection for Mold: The proposed dust sampling method for mold is a vacuum (fitted with a fine nylon mesh tubular fabric sampler inserted in the vacuum hose wand). The perimeter of the rooms (and if necessary underneath the furniture) will be vacuumed for floor dust samples. The mold dust samples are intended to be analyzed for fungal species found in water damaged environments using PCR analysis and a patented EPA DNA extraction method.

Building Moisture: Building moisture measurements will be taken using a digital, pinless moisture meter. When placed against a surface, these instruments provide a reading that is proportional to the moisture content of the surface. Measurements are to be taken at three distances from the floor; 3 inches, 3 feet, and 6 feet.

Surface Wipe Sampling for Pesticides: Pesticide residues will be measured by collecting surface floor wipes (using EPA's *Analytical Protocol (AP) for the Collection of Surface Wipe Samples (i.e., Tabletops, Flooring (Hard Flooring, Carpeting)), HEAB-AP-20.02.1-HUD, Revision 1, dated February 20, 2001*) from two rooms where the respondent reports that pesticides have been applied in the last 12 months. If no pesticide application is reported, the randomly selected common living area and bedroom will be sampled. The one square foot samples will be collected from along the wall.

Air Sampling for Formaldehyde: For the AHHS survey pilot study, a passive diffusion badge was used for formaldehyde sampling. However, EPA may select a different device to meet its sensitivity needs. If the badge is used, formaldehyde samples will be collected from the randomly selected common living area and bedroom. Badges will be hung in the rooms at the beginning of the visit and remain until the end of the visit. The samples will be sent to a lab for analysis using High Performance Liquid Chromatography.

Carbon Monoxide Emission Testing of Appliances: The specific procedures for measuring carbon monoxide emissions and the conditions under which the measurements are taken have yet to be determined. In the pilot study for the survey, a direct reading instrument was used.

Sampling for CCA (Arsenic): The proposed sampling plan for CCA includes both a floor wipe sample inside the main entrance and a soil sample. CCA soil samples will be collected in accordance with EPA's Draft Protocol for Sampling for Residues of Arsenic, Chromium, and Copper in substrates (soils/buffering materials) Beneath/Adjacent to Chromated Copper Arsenate Treated Playground Equipment, dated 9/18/01). The top one inch of soil will be collected from an area two inches from the wooden vertical support. The soil and wipe samples will be sent to a laboratory for arsenic analysis.

Vacuum bag/Dust Analyses: EPA will specify the requirements for analyses that may be conducted on the vacuum bag dust. These may include nicotine, perfluorooctanoic acid (PFOA), and bioavailable arsenic.

Vermiculite Attic Insulation: The technicians will be trained to identify different types of attic insulation. These include vermiculite which may contain asbestos. If the attic insulation is accessible, the technicians will record the type of insulation and take pictures of the insulation. To avoid asbestos hazards, the insulation will not be sampled or otherwise tested for asbestos.

Quality Assurance/Quality Control: The principal source of quality assurance will be the utilization of well-planned, detailed and tested documented protocols for all aspects of data collection: listing, in-person interviews, field observation, measuring and recording physical data, collecting environmental samples, equipment and sample handling, and data management. Thorough study-specific training of experienced field staff will be critical to assurance of a quality product. Finally, ongoing communication between and among the various individuals responsible for each stage of the study will be rigorously maintained to assure quality.

The quality control for field data will include:

- Manual edit of data collection forms by field staff;
- Review of data collection forms by the field supervisor;
- Telephone verification of data collection for a random 10 percent of the households;
- Random field audits;
- Quality control samples and measurements

Field quality control samples for lead measurements are shown in the following table. Replicate lead measurements will not be collected in AHHS because the NSLAH survey has adequate data for assessing components of variance. Quality control samples and measurements for other analytes have yet to be determined.

Table 1. Quality Filed Control Measurements/Samples for Lead

Purpose	XRF	Dust Wipe	Soil
Materials Screen	NA	2 per lot of supplies	2 per lot of supplies
Field Blank	XRF calibration check– 2 per HU (pre/post)	Not Applicable	NA
Reference Sample	XRF Calibration of SRM film – 2 per HU (pre/post)	Made from NIST reference soil – 30 high 30 low per lab	NIST soil 30 high plus 30 low per lab
Sample Replicate	XRF Calibration – 3 SRM measurements – 2/HU (pre/post)		

Notes: HU = Housing unit; SRM = NIST Standard Reference Material

Quality Control for Data Management, Preparation and Analysis: All data requiring key entry will be re-typed 100 percent. Any discrepancies will be immediately resolved. The resultant data and test results will be submitted to computerized range and logic checks. All discrepancies and out-of-range values will be investigated and resolved.

Compiling the survey data files from various sources requires multiple processing steps. The results of each processing step will be checked for accuracy. In addition, all data from 2 percent of homes will be printed and manually compared to the original data sources to check that the final data files accurately reflect the original data.

Finally, the data files will be formatted to make the data structure and data values easy to understand. The data documentation will describe each file, each variable, and value in each variable as well as provide guidelines for using the data.

Statistical Issues: To fully account for the complex survey design, it is necessary to apply sampling weights to each completed case. A given unit's sampling weight is roughly the number of housing units nationwide represented by the study unit. The initial weights may be further adjusted to balance differences in nonresponse and noncoverage. The sample weights will be available for analysis and variance estimation.

In some cases the precision or bias of statistical estimates depends on the variance components, for example, differences between rooms in a home or differences between surfaces in a room. The magnitude of the variance components can be assessed from replicate samples, such as samples from difference surfaces in the same room. For some contaminants, replicate samples will be taken to estimate the variance components.

Homes can be classified as having or not having any lead-based paint. However, the estimated proportion of homes with LBP can be biased by a combination of 1) within surface variation and measurement variation, and 2) incomplete sampling of rooms. For the NSLAH survey the bias was estimated based on the variance components. Similar analyses may be necessary for the AHHS results.

The precision of the survey estimates depends on the estimate under consideration and the subpopulation of interest. The sample design and sample size is based on a combination of cost and precision considerations. The approximate width of a 95 percent confidence interval for a percentage estimate for all homes is expected to be about +/- 3 percent, with correspondingly wider confidence intervals for smaller subpopulations.

Concluding Remarks: The AHHS survey will provide valuable information about hazards in home environments. An important objective of the survey is to provide data for comparisons across time. Since comparable household data for lead hazards and allergens was collected in the NSLAH survey, changes across time can be estimated for these measures. In addition, XRF paint measurements were collected in the 1990 survey of lead paint. These measurements provide another point of comparison. However, a different XRF instrument was used in that survey.

The other analytes to be measured in the AHHS survey will provide a baseline against which future data collection efforts can be compared. Rather than a survey of a local area or hotspot, the AHHS is a national survey. The AHHS results can provide national estimates against which the results from local or hotspot studies can be compared.

Pesticide Epidemiology, Biomonitoring, and Risk Assessment: Four Case Examples

Allen, RH* Christensen, CH* Conomos, MG ** and Blondell, J*

*US EPA, 7509-C Office of Pesticide Programs, Health Effects Division
**Office of Environmental Information, Environmental Analysis Division

EPA Quality Meeting, April 15, 2004, Tampa Fla.

Introduction

The objectives are briefly to: (1) summarize four case examples of current pesticide epidemiology, biomonitoring, and risk assessment and challenges they raise, and to (2) highlight areas for the development of web-based education for staff risk assessors, outreach to stakeholders and the general public, plus training for public health professional.

Methods

The four case examples are part of ongoing analyses by the Office of Pesticide Programs, Chemistry and Exposure Branch, Epidemiology Group in support of legally mandated pesticide risk assessments. These assessments comply with the Federal Food Quality Protection Act (FQPA) of 1996, earlier requirements of the Federal Insecticide Fungicide, Rodenticide Act of 1988, and guidance on risk assessment and technical papers on aggregate and cumulative exposure assessment. Data sources are:

(1) Pesticide poisoning incident reports from multiple sources,
(2) Pesticide epidemiology findings from the inter-agency Agricultural Health Study (AHS),
(3) Pesticide biomonitoring findings of the various National Health and Nutrition Examination Survey (NHANES) reports
(4) Pesticide epidemiology reports in the open scientific literature.

The four case examples are being examined here as part of a larger strategic planning effort to design better education, outreach and training in pesticide epidemiology. There is a need for clear and concise web-based information, in sync with other environmental information and analysis

Results

Example #1 Pesticide poisoning incident reports

The general perception is that pesticide poisoning incident reports are nice to have, but "we do not regulate on them," but this is a misperception. Initially, this was true when the only data available were sporadically published clinical case reports of adverse health effects from unexpected exposures or intentional suicides. Many of the adverse effects were transient or acute, rashes, eye irritation allergy, hypersensitivity, and respiratory effects. Currently, there are five main tracking sources of pesticide poisoning information, and US GAO has twice rated reporting as incomplete and seriously flawed. A training program has been created to train health care professionals in the recognition and management of Pesticide Poisonings, and EPA contracts for the periodic publication of a manual on the same topic, as a reference service for

health professionals. The first data source is the 2000 cases reported by industry to EPA under FIFRA 6 (a) 2 Incident Data System. The second data source is the 110,000 reports per year to Poison Control Centers, with an additional 27,500 cases reported to poison control centers and seen in health care facilities, including 25,700 cases that have poisoning related symptoms (84% minor, 15% moderate, and 1% major or fatal). The third data source is the 1200 reports per year to the California EPA equivalent. The fourth data source is the 500 State reports to the growing NIOSH SENSOR program. The fifth data source is the 900 reports per year to the National Pesticide Information Center, maintained by EPA under contract to Oregon State University.

The two key information needs are education for prevention, including the training of health professionals, and improvements to surveillance tracking. In over a dozen instances, pesticide epidemiology information on poisoning incidents was used for risk recommendations.

Example #2 Pesticide epidemiology findings from the inter-agency Agricultural Health Study (AHS)

A number of recent publications from the inter-agency Agricultural Health Study (AHS) suggest where pesticides are not or are associated with chronic health effects, e.g., macular degeneration, prostate cancer, or lymphohematopoietic cancers. The details of these reports are beyond the scope of this report, and can be found at the agricultural health study web site listed in the Web Resources. The eye disease findings sparked follow-up field studies, since back sprayers of fungicides had elevated risk, as did those with vision impairment. For the prostate cancer findings, a project is now underway to examine the pesticidal inert ingredient, many of which are not biologically inert. Also, findings from the prostate cancer paper showing an excess risk of 2 fold for use of 7 of 50 and 4 fold excess risk for those with a family history and ever/never use of the compounds are likely to prompt toxicological studies and studies of susceptibility genes tested in genetically engineered rodents.

The information needs that emerged from the findings so far are: (1) non-specialists need a primer to understand the basic concepts of epidemiology before beginning to grapple with the question “What do the numbers mean?” A publication on pesticide epidemiology from Purdue University has been helpful. (2) There is need for additional guidance and web-based resources on the interpretation of pesticide epidemiology findings for regulatory purposes. This need is growing due to the rapid increase in quality and quantity of publicly available pesticide epidemiology information. (3) Cutting-edge science or gene-environmental interactions, toxicogenomics are needed based on the observation that men with a family history of prostate cancer have a doubling again of risk that double in association with 7 of 50 pesticides, and (4) Pesticidal inerts with potential cancer producing properties may need to be re-examined for “prevention opportunities.”

Example #3 Pesticide biomonitoring findings of National Health and Nutrition Examination Survey (NHANES)

The NHANES pesticide biomonitoring findings provide a rich new source of pesticide epidemiology information. Ongoing analyses are of two types: chemical specific assessments for inclusion in reregistration eligibility documents (RED) posted on the EPA web site, and methods development including risk model validation for multiple routes of exposure. Twelve specific analytes are available of NHANES III, and over 30 are available for NHANES ‘99-‘00, with two

years of additional data expected shortly from NHANES ‘01-‘02. Earlier pesticide measurements from the Hispanic HHANES of ‘82-‘84 are comparable for 6 analytes and these results are being evaluated for development of results indicators.

The information needs that emerged from these assessments are: (1) better characterization of the role of creatinine correct for both adults and children, (2) a wider array of analytes, since many compounds are used widely but relatively few are part of regular biomonitoring, and (3) environmental statistical weighting issues for complex samples as challenge to communicate to diverse audiences of varying statistical literacy.

Example #4 Pesticide epidemiology reports in the open scientific literature
In the last year, a number of published pesticide epidemiology studies of different designs reached senior management from different sources and required explanations of “what do the numbers mean” which in turn requires an understanding of study design and population characteristics. There are a chain of disconnects. Ecological studies may suggest hypotheses that go without follow-up, case control studies are often too small or geographically limited, thus putting even greater reliance of the emerging data from well designed and executed, but rarer prospective studies with data over time, but largely lacking biomonitoring. At the same time, the National Academy of Science concurred on the controversial recommendation for allowing studies of intentionally dosed humans to be used in regulatory decisions, subject to some as yet unspecified, and difficult to specific human subjects protection testing guidelines. So, human data now has a double meaning, for an audience with very uneven depth of understanding in the basic concepts.
The information needs from this situation are: (1) pesticide epidemiology education at all levels, (2) focused staff training in pesticide epidemiology, especially for risk assessor, and (3) web-based training in basic pesticide epidemiology for diverse external audiences. Such efforts are underway this fiscal year.

Conclusions
Pesticide epidemiology, biomonitoring and risk assessment have many embedded statistical, information and communication challenges that need to be addressed if the information is to serve the rigors of regulatory decision making.

Web Resources
http://www.epa.gov/opppsps1/fqpa/
http://www.aapcc.org
http://search.epa.gov/s97is.vts
http://www.cdc.gov/nchs/nhanes.htm
http://www.aghealth.org
http://ehp.niehs.nih.gov/

Modeling Hazard Waste Arrival and Single Server Incinerator In Fixed Time: Monte-Carlo Approach

Nelson Andrews; Economics, Methods and Risk Assessment Division (EMRAD) of the Office of Solid Waste (OSW), U.S. EPA

The modeling of the formation of possibly toxic clouds, as hazardous waste is burned, requires the examination of the rate at which hazardous waste arrives at the burn site and the rate at which the waste is processed and incinerated. From a statistical perspective, this becomes a stochastic process with a queue and a single server (incinerator with a finite capacity). To simplify the mathematics for the derivation of the model, I assume an arrival rate for a fixed unit of hazardous waste and assume that this is the same unit of hazardous waste that is processed and incinerated each time the incinerator is used for burning hazardous waste. Other assumptions used are (1) the independence of the arrival times of hazardous waste and the lengths of time for incineration of hazardous waste upon arrival; (2) arrival times form a Poisson process and (3) the length of incineration times form a Poisson process. With these assumptions, it can be shown that this arrival/service process, where the state is the number of units of hazardous waste in the system waiting for service or being served, is a Markov process (has short term memory). This problem, even with the Markov process assumption, is not an easy problem to solve, but using a Monte-Carlo approach, we can use random number generators to simulate the behavior of such a queue and server for estimating the respective probabilities

As I began researching various modeling efforts related to hazardous air pollutant emissions, it became apparent that there are as many models as there are different terrains in the environment. To maintain my sanity (and not wishing to turn a short term fact finding mission into a thesis), I quickly narrowed the field to the EPA Preferred models. I briefly mention several of these models to support the fact that source emission rate (Q) is a critical input factor for such models. This paper looks at the business related parameters of a plant (source) such as hazardous waste arrival rate and the material processing rate to achieve a source emission rate over a fixed time period. I assume that the incoming material is hazardous waste to justify the social and economic benefits to producing of a toxic plume. It is conceivable that the incoming material is not hazardous but merely converted during a production process or an Energy recovery process.

Quick Look At EPA Preferred Models

The preferred models that I mentioned earlier can be found on the EPA "modeling" web page http://www.epa.gov/scram001/tt22.htm . On this web page, there are six models described. Of these six models, four are appropriate for terrain stationary sources. The four models are:

1) Buoyant Line and Point Source Model (BLP),
2) Complex Terrain Dispersion Model Plus Algorithms for Unstable Situations (CTDMPLUS),
3) Industrial Source Complex Model (ISC3),
4) Offshore and Coastal Dispersion Model (OCD).

In each of these four models, a required input is the source emission rate for the hazardous air pollutant plumes being emitted from the stack(s).

(BLP Model)
The BLP dispersion model models the point source and the line source plume rise and physical distribution over a given terrain. The model uses hourly meteorological data such as wind speed, wind direction, ambient air temperature and mixing height to model the dispersive characteristics of the atmosphere. Restrictions for the model are a maximum of 50 point sources and a maximum of 10 line sources. Each line source is assumed to have equally spaced point sources with stack parameters such as height, width and base elevation are constant within each line. In general the model categorizes the model either as rural or urban.
For the pollutant emission rate (g/s), whether it's the point source emission rate (Q) or the line source emission rate (Q_L), the model assumes a single average emission value for the time under study. Since there is a maximum of 366 days for the meteorological input, I assume that the maximum time period that can be considered for this model is one year. For the point source input, the emission rate may vary from point source to point source. For the line source input, the emission rate may vary from line source to line source. But for the point sources within the line source, the emission rate is constant.

(CTDMPLUS)
As the name suggests the CTDMPLUS dispersion model is designed to address the complex terrains. The inputs for this model are similar to the previous models with the exception that a mathematical depiction of the shape of the terrain is required as input. When using this model the user first describes the terrain then provide the meteorological data and the emission source data. Unlike the previous model (BLP), this model deals only with point sources (no line sources). By default it can model up to 40 sources, but the user may change the maximum number of sources.

The pollutant emission rates (mass per second – g/s) may be given in terms of an average value per source or the user has the flexibility to enter one emission rate per hour per source. The second choice gives the user more flexibility in describing the pollutant emission rates over time when both stable and unstable conditions are present.

(OCD Model)
The OCD dispersion model describes the onshore impact of offshore pollutant emission sources. This model uses hourly onshore and offshore meteorological data to model the behavior of plume rise and dispersion. Since offshore source stacks may not be at a zero degree angle from the vertical, this model introduces an additional parameter; the stack angle. In addition to using either point sources or line sources, this model makes provision for area sources. The area source is designated as a circular area with center located at the Cartesian point (x,y).

The source input for this model is similar to the previous models. The pollutant emission rate (Q) for the source is in terms of mass per second. The input may be constant for each source or it may be hourly emission data may. It is not clear to me how many point sources can be modeled and the maximum time period. I will leave this for future investigation.

(ISC3 Model)
In the industrial complex environment, the ISC3 dispersion model includes the layout of the terrain grid to account for the dry deposition of the generated plume along its path. Like the previous models, it also divides input into meteorological and source parameter data. The terrain grid information is the third input component for this model. The surrounding terrain for this model is one of two choices; either rural or urban.

The source input for this model requires at least a constant pollutant emission rate throughout the modeling period for each point source. Unlike the other models, this model only considers point sources (a maximum of 500). The point sources may be grouped into 2 to 50 groups. However, this model does handle a variety of types of sources. A source may be one of the following: Point (A single point source such as a stack), AREA (ground level releases such as lagoons or dumps), VOLUME (multiple levels such as buildings and vents) and OPENPIT (ie. Rock quarries). For each source, the emission may range from an hourly emission rate (for the short term model only) or a value that is constant over the modeling period.

Emission Rates and Business Operation Parameters
After this brief and painless overview of the 'EPA Preferred' Models, it is apparent that the point source emission rate is a required and necessary input value when implementing plume dispersion models. So, how is the average pollutant emission rate computed for a typical point source? In most instances, the emission rate for a source is measured for a specified length of time and the average value of the empirical measurements becomes the average emission rate for the source. This is good, when the business operating parameters will not change, but how realistic is such an assumption?

So, let's get to the business of relating the business operating parameters to the pollutant emission rate of a point source. The two business operating parameters that are relevant to the emission rate are (1) the delivery rate of the raw material and (2) the service rate of the raw material (In the case of Hazardous Waste Incinerators, this is the rate at which the raw material is processed and incinerated.). To simplify the calculations that follow, let's assume that the delivery occurrence is a poisson process with expected inter-arrival time of λ. For the service occurrence let's assume a poisson process with expected processing time of μ. The final parameter that is required is the conversion factor between the raw material and the hazardous emission pollutant. We will call this factor 'ρ' and it relates the amount of raw material processed ($\mathbf{X}$) to the amount (mass) of pollutant produced ($\mathbf{Y}$) as follows: $\mathbf{Y} = \rho(\mathbf{X})$.

If the average amount of raw material processed in the fixed time $\mathbf{T}$ is denoted as $\mathbf{X_T}$, then using the conversion factor we have the amount of pollutant produced in this time is $\mathbf{Y_T} = \rho(\mathbf{X_T})$. Using the notation for average emission rate for the pollutant of interest, we denote the average emission rate over time $\mathbf{T}$ as $\mathbf{Q_T}$. Dividing the amount of average amount of pollutant produced over time $\mathbf{T}$, by the time $\mathbf{T}$, we have the following formula for the average pollutant emission rate:

$$\mathbf{(1)}\ \mathbf{Q_T} = \rho(\mathbf{X_T})/\mathbf{T}.$$

(Use the appropriate multipliers to convert the resultant number into g/s units.) Now we have a formula (**1**) that takes us from the raw material average processing rate to the average pollutant

emission rate. If we have no raw material to process, we have no pollutants being emitted. Therefore, the amount processed must be related to the amount of material arriving at the point source. How do we relate the arrival rate to the amount processed? The next section discusses this relationship while setting up a stochastic process to relate these parameters.

Queuing Theory Approach for Plume Production

Examining a basic unit (fixed amount per delivery) of raw material for delivery and incineration, we can treat the number of units remaining in the queue after an elapsed time T as a birth/death process. Making the assumptions discussed in previous section about the distributions of the arrival rate and the processing rate, we can simplify the mathematics required to determine the rate of processing each unit of raw material. But even with these simplifying assumptions, trying to build a mathematical expression becomes a significant challenge in algebraic manipulation. Hence, later we will use a Monte Carlo approach to estimate the probabilities of '$\mathbf{K_T}$' remaining when we assume Poisson arrival process (Markov) and a Poisson service process (Markov). In this instance we will assume a single incinerator for each source (hence a single server). From queuing theory notation, we have the following shorthand notation **M/M/1/∞**. The infinity symbol, '∞', indicates the queue size is significantly large enough to hold any number of units that arrive.

As we use the properties of the **M/M/1/∞** queue to construct probability functions, we need to define some additional variables for this arrival/service process. We have already introduced '$\mathbf{K_T}$', the remaining units in the queue. Now we build a list of all the pertinent variables for this queuing problem in the table below.

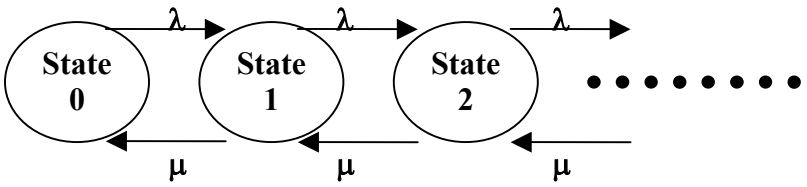
Table 1: **Stochastic Variables for analysis of Pollutant Emission Rate**

Stochastic Variable	Description	Comments
$\mathbf{K_T}$	Number of units of raw material remaining in queue after elapsed time 'T'.	(Assume queue empty at time zero.)
$\mathbf{N_T}$	Number of units of raw material processed during elapsed time 'T'.	
$\mathbf{A_T}$	Number of units of raw material processed during elapsed time 'T'.	$\mathbf{A_T = K_T + N_T}$
$\mathbf{C_F}$	Conversion factor for converting units of raw material to mass of raw material.	$\mathbf{X_T = C_F\ E(N_T)}$, The notation **E( )** is used to represent expected value.
$\mathbf{t_1, t_2, \ldots, t_n}$	Sequence of stochastic independent delivery arrival time intervals. (The I^{th} arrival will occur $\mathbf{t_I}$ length of time after $(I-1)^{th}$ arrival.)	
$\mathbf{s_1, s_2, \ldots, s_n}$	Sequence of stochastic independent service times.	

Next we investigate the steady state behavior of this queuing model. The steady state refers to the stochastic behavior of the model as the total elapsed time approaches an infinite value. To derive the stochastic behavior (probability of being in a given state after a substantially large

elapsed time period), we construct the state model. The state model for the **M/M/1/∞** queue, when the state is the number of units remaining in system, is shown in Figure 1 below.

Figure 1: State Diagram Model for **M/M/1/∞**



The parameters μ and λ in Figure 1 refer to the expected service time and the expected delivery time respectively. Establishing the steady state equations for this model and solving the equations for the probability of being in state '**k**', we get the following formula:
Probability {State k} = $(\mu/\lambda)^k (1 - \mu/\lambda)$.
In reality, this steady state time period happens after an extended elapsed period of time; but whether a source is consistent enough in their operations to reach steady state for given parameters is purely subjective. However, the steady state values provide a check for the accuracy or correctness of the results obtained through the Monte-Carlo approach. By projecting the results from the Monte-Carlo approach out to an infinite time period, we should get results that are close to the steady state probabilities. This method is shown in the next section.

Monte-Carlo Approach

The heart of the Monte-Carlo approach is a good Uniform(0,1) random number generator. But the random number generators on most Personal computers probably fall in the area of pseudo-random number generators. For the precision of this modeling effort, such generators are adequate. The approach for this effort is the following:

1. Use the random number generator to generate a random number between 0 and 1. Such numbers will be generated throughout this methodology and will be used to obtain a corresponding random variable associated with some distribution function of interest. Let '**p**' represent a random generated number between zero and one. If **F(x)** is a cumulative distribution function, then $\mathbf{F^{-1}(p)}$ is a random variable with cumulative distribution function **F(x)**.
2. Given the fact that the arrival times form a Poisson process, the random number of arrivals in time '**T**' is distributed according to a Poisson distribution with an expected value of $\mathbf{T/\lambda}$. Using $\mathbf{P(}A\ ;\ \mathbf{T/\lambda}\ )$ to represent the cumulative probability of less than or equal to k arrivals in time '**T**' with mean $\mathbf{T/\lambda}$, we generate the random number of arrivals in time 'T' with the following: $\mathbf{A_T = P^{-1}(p;\ T/\lambda).}$
3. After generating the random number of arrivals ($\mathbf{A_T}$) in time **T**, we now derive the time of occurrence for the $\mathbf{m^{th}}$ delivery. We begin this task by letting $\mathbf{T_m}$ represent the time of the $\mathbf{m^{th}}$ delivery or $\mathbf{T_m = t_1 + ... + t_m}$. From the properties of the Poisson process, the

conditional probability of $\mathbf{T_m}$ being less than τ given that there are exactly $\mathbf{A_T}$ arrivals in time **T** and the $(\mathbf{m-1})^{th}$ time of arrival is $\mathbf{T_{(m-1)}}$ is computed to be $\mathbf{1} - [(\mathbf{T} - \tau)/(\tau - \mathbf{T}_{(m-1)})]^{\mathbf{A(T)-m+1}}$; where $\mathbf{A(T) = A_T}$. Equating this function to the uniform random variable '**p**' for the distribution of each $\mathbf{T_m}$, we can generate the random delivery times for all $\mathbf{A_T}$ delivery times in elapsed time **T**.

4. By definition of the service times forming a Poisson process, each random variable $\mathbf{s_i}$ (i = 1, …., $\mathbf{A_T}$) is randomly distributed under an exponential distribution with expected value μ. Using this fact we generate each of the random service times for $\mathbf{s_i}$ by equating the cumulative distribution function for each service time to a random value 'p' and solve for the service time.
5. The final step in the Monte-Carlo approach is to determine the number of units that complete service during elapsed time **T**. To derive this value, we construct a sequence $\mathbf{S_k} = \mathbf{max(T_k , S_{(k-1)})} + \mathbf{s_k}$. Then we search for the largest value '**k**' such that $\mathbf{S_k} < \mathbf{T}$. This value gives us the stochastic variable $\mathbf{K_T}$ for this particular run of the Monte-Carlo approach.
6. Repeating steps 1 through 5 above a large number of times, we construct a frequency table and provide estimates for the probability of being in a given state after a fixed elapsed time '**T**'.

From the frequency distribution table, generated through the Monte-Carlo approach, we are able to construct the expected number remaining in the system at elapsed time **T**. Using the relationship $\mathbf{A_T} = \mathbf{K_{T+}}\ \mathbf{N_T}$ and the fact that the expected value of $\mathbf{A_T}$ is $\mathbf{T}/\lambda$ we have the following formula for the expected number of units processed in time **T**: $E(\mathbf{N_T}) = \mathbf{T}/\lambda - E(\mathbf{K_T})$. With this expected value of the number of units processed in time **T**, we are now able to compute the expected amount processed in time **T** as follows: $\mathbf{X_T} = \mathbf{C_F}\ \mathbf{E}(\mathbf{N_T})$. Upon reaching this stage in the analysis, we can call it a day and say a job well done. But before, we quit I will give a few thoughts about the verification of the accuracy of this approach.

In theory, when the steady state probabilities can be computed, the fixed time probabilities should approach the steady state probabilities as we evaluate the fixed time probabilities for larger values of **T**. To verify the accuracy of the Monte-Carlo approach, I fit the results for the Monte-Carlo probabilities to a geometric distribution and provide a parametric distribution function for the fixed time probabilities. For a fixed state, I extrapolate the value of the fixed state probability as **T** becomes large (It should be near the steady state value.). Now we can call it a day.

Region 5 Changes In Estimated Hazard Exposure and Demographic Characteristics: 1990-2000

by
Lawrence Lehrman and Arthur N. Lubin

Introduction

There were several procedures used to show areas of migration of population and changes in toxicity hazard estimates between 1990 and 2000. The changes in hazard estimates between 1990 and 2000 were compared among levels of demographic attributes. The demographic characteristics were population densities, proportions low income, proportions of resident populations who are members of minority population categories and numbers of new residences built during the ten year period. The estimated hazard exposure differentials among the demographic characteristics indicate whether or not populations with different characteristics also experienced different changes in estimated toxicity exposures. Changes in hazard estimates were compared with changes in population sizes. This certainly has policy implications because of the EPA's Environmental Justice efforts designed to ensure environmental "fairness." Furthermore, changes in hazard estimates were shown. Finally, discriminant analysis was done to estimate whether or not the population size change quartiles provided maximally homogeneous areas in terms of selected demographics and the estimated 2000 hazard estimate.

Data-base Development

The initial stage of the effort involved the development of a data base which combined EPA approved risk coefficient values, TRI data along and 1990 and 2000 U.S. Census block group level information. Sources of the data and coefficients were the EPA's TRI data base, the Bureau of Census 1990 and 2000 and EPA's Risk Screening Environmental Indicators (RSEI) also from 1990 and 2000.

The RSEI was produced by the Economics, Exposure, and Technology Division Office of Pollution Prevention and Toxics. RSEI does not calculate detailed or quantitative risk assessment but offers a screening-level, risk-related perspective for relative comparisons of chemical releases. The risk-related score is a unitless value are proportional to the potential risk-related impact of each element. Actual scores per TRI facility are derived by summing the estimated risks from each element emitted. Although the model results do not capture all environmental releases of concern, they do relate changes in releases to relative changes in chronic human health impacts from a large number of toxic chemicals of concern to the Agency. Our purpose for using RSEI was to compare relative risk levels over time and definitely should not be viewed as an effort to predict actual risk levels.

U.S. Census block-group level data were used to estimate the changes in population densities over time (1990 - 2000). The primary reason block-group level data were used is that it is the smallest geographic area that incorporates sampled data. We required sample data because our

analysis includes the census parameters of new homes (built since the last census) , minority populations and low income.

Cell grids are an ideal tool for analysis due to spatial uniformity (census blocks groups were inconsistent in the 1990 versus 2000 censuses). A uniform grid cell system is excellent for incorporating data from polygon features with inconsistent boundaries. It also produces a continuous and uniform product allowing us to complete boolean queries, algebraic functions, and statistical analysis from one table. We used a 3 X 3 km cell size to be consistent with the resolution of block groups in the suburban areas experiencing growth.

A surface allocation technique was used for the purpose of smoothing small spatial variations in the demographics data. Our technique involved having the demographics used to populate the cell be representative of a larger area than the cells area. Our 3 by 3 km grid cells represent the demographics of the centroid buffered to 5k. The buffered polygons are intersected with the block-group Census polygons and the populations estimated. The hazard allocation weights for grid cells were calculated using much the same procedure as for the demographics data. The grid cell RSEI attribute consists of the sum of the hazards within the 5k buffer. A 5k buffer provides for demographics and releases from facilities near the edge of the cell to be considered as having an influence.

The Region 5 grid consists of 99,639 3 by 3 km grid cells. The attributes are the 1990 and 2000 Census Demographics (population density, new homes built since last census, and the Environmental Justice parameters of minority, low income and poverty levels) and the 1990 and 2000 RSEI hazard densities. Population and toxicity changes over time were calculated by subtracting 1990 from 2000 results. The population and hazard changes were quartiled and the results were mapped.

Analysis and Results

An initial objective in this analysis is to determine areas of apparent increasing risk using the RSEI Hazard numbers and comparing them across demographic attributes. An advantage of a grid system is it makes querying an easy procedure. The attributes in our grid table includes Grid-id and 1990 and 2000 population and hazard densities. Differences between the 1990 and 2000 population sizes and hazard estimates were calculated and quartiled.

The statistical analysis involved obtaining averages for the demographic change variables (2000 minus 1990 values) and the estimate of risk in 2000. This was done for all of the approximately 100,000 squares aggregated as well as per demographic quartile (the quartiles of population size 2000 minus population size 1990 were used as groupings). The averages per quartile of the risk estimates and the demographic variables are shown on Table 1.

.The quartile means of the average hazard estimate indicate the expected pattern of persons tending to migrate into are as with lower exposures. It is interesting that the differences among the quartiles in terms of most of the demographics do not seem substantial.

Table 1

Average Level of Hazard Estimate and Demographic Change Variables (2000 minus 1990 values) Per Quartile of Population Change Variable (2000 minus 1990 population)

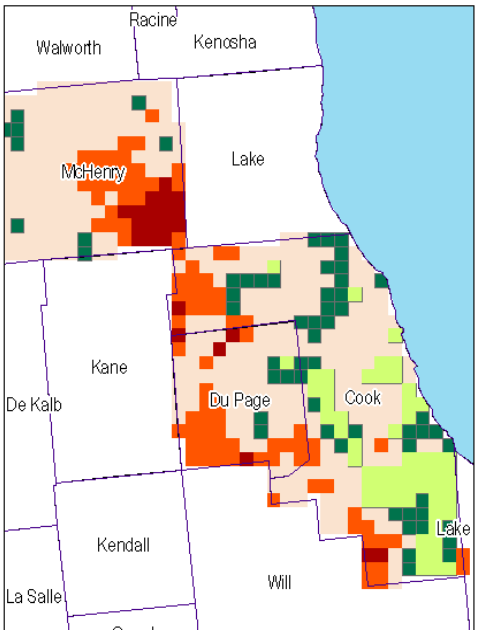
Quartile of Population Change	Average Hazard Estimate	Average Change in Population Density	Average Change In Prop. Low Income.
Quartiles Aggregated	184,440	12.60	-2.97
1	395,286	-8.47	-2.60
2	161,018	2.71	-3.11
3	109,520	13.28	-3.25
4	71,336	43.06	-2.93

Quartile of Population Change	Average Change In Prop. Minority.	Average Prop. New Housing (post 1990)
Quartiles Aggregated	3.52	2.50
1	3.80	1.41
2	3.40	2.01
3	3.10	2.87
4	3.78	3.71

The apparent relationships between the population change quartiles and the hazard 2000 quartiles were further illustrated using the map queries in the legend for Figure 1. .

Figure 1

The areas which are not colored per the legend for Figure 1 are residual areas where the population growth and hazard growth estimates are in opposing quartiles and/or- halves. For instance, the area may be in the upper quartile or half for the hazard change estimate and in the lower quartile or half for population growth. Because a high proportion of the Chicago PMSA is not colored per the legend, Figure 1 further suggests an aversion to migrating into areas with higher hazard estimates.



The Ever Changing and Expanding Universe of

TRI and RSEI should be considered when interpreting Figures 2 and 3 which follow. The TRI program has expanded significantly since its inception in 1987. The Agency has issued rules to roughly double the number of chemicals included in the TRI and seven new industry sectors have been added to expand coverage significantly beyond the original covered industries. Most recently, the Agency has reduced the reporting thresholds for certain persistent, bioaccumulative, and toxic (PBT) chemicals .Overall in the Region 5 Urban areas there was a 17% percent reduction in total estimated hazard from 1990 to 2000...

Figure 2 is a map of much of Region 5 indicating areas with increasing or decreasing RSEI numbers. The results mapped in Figure 2 were obtained by quartiling the cell's hazard exposure estimates for 1990 and 2000 separately and subsequently subtracting the 1990 from the 2000 quartile for each cell. Figures 3 shows similar results for the Chicago PMSA. Positive results from indicate areas where relative risk increased and negative results where relative risk decreased.

Figure 2

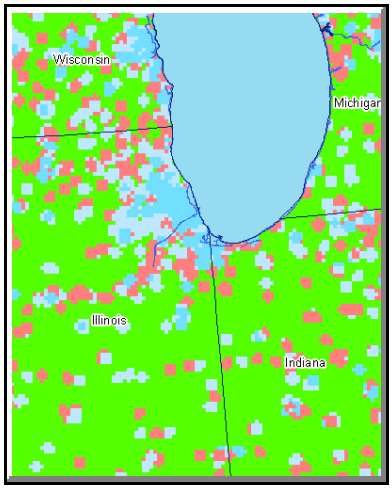
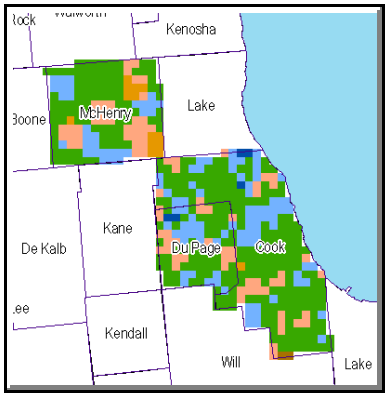
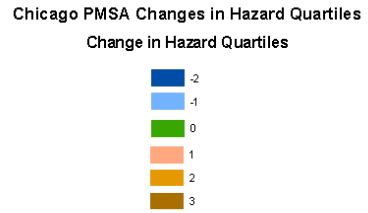


Figure 3





Another interest was whether or not the population change quartiles provided relatively homogeneous areas in terms of the demographics and the estimated 2000 hazard toxicity exposure. This was tested using discriminant analysis. Discriminant analysis uses the generalized squared distances (based upon the hazard and demographic values) among the block groups to categorize the groups into categories while taking into account the original categorization probabilities. The pooled covariance matrix was used because the within group covariance matrices were relatively equal. Version 8.0 of the Statistical Analysis System (SAS) Software was used to perform the statistical analyses.

The results of the discriminant analysis categorizations versus those for the original quartiles for the entire Region provided a fairly high level of categorization agreement. Approximately 65 per cent of the cells's categorizations were not altered by the statistical approach. Exemplary results with a similar level of categorization agreement are provided by the maps of the Detroit urban area shown below. Figure 4 shows the original quartile groupings and Figure 5 the discriminant groupings. In addition, it was relatively unusual for the discriminant analysis to yield categorization changes which reallocated a cell to more than one grouping level different from the original result. This appears to suggest that the relatively simplistic quartile based grouping approach is not necessarily woefully inadequate for creating at least somewhat homogeneous categories. The creation of relatively homogeneous areal groupings is potentially valuable for a wide range of applications including developing more efficient sampling strategies and possibly for the targeting of area-wide environmental impacts mitigation efforts.

Figure 4

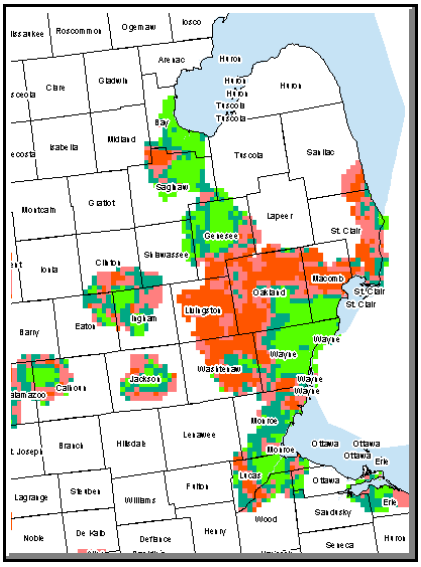
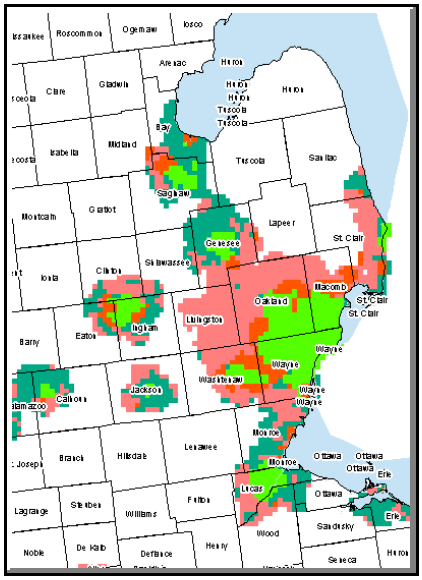
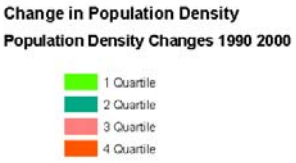


Figure 5





Summary and Results

The effort has provided several significant results. First, the study found that in Region 5 the areas with greater TRI related hazard estimates tend to have reduced population growth. However, the relationships between risk estimates and changes in demographic characteristics do

not appear to be substantial. Thus, the results do not seem be indicative of increasing levels of environmental inequity. The finding that higher risk estimates may be related to lower population growth was further verified by Figure 1 which demonstrated that there are substantial areas within the Region where the risk is decreasing; especially those areas with reduced population growth. Second,the relative success of EPA's hazard reductions programs is suggested

by Figures 2 and 3 which demonstrated that a high proportion of the map cells had lower relative risk estimates in 2000 versus 1990. The maps probably understate the success of the program due to several areas having artificially increased risk levels in 2000 versus 1990 at least partially due to the expansion of the number of chemicals included in TRI, new industries being added to the data base and reduced reporting thresholds for several chemicals. Third, the discriminant analysis demonstrated that the relatively simplistic approach of quartiling based on population growth provides at least somewhat homogeneous groupings in terms of 2000 risk estimates and the selected demographic characteristics.

The present effort should be viewed as merely a work in progress. There are several additional areas where future efforts probably should be directed. First, if the investigation focused on additional areas (smaller or larger levels of scale), would the results be similar to the present findings. In other words, are the results geographically generalize able? Second, the specific characteristics which account for the relative consistency of groupings is uncertain. This could be determined via variance decomposition approaches. Third, similar investigations should be done to determine the findings which would be obtained if the analysis was done for a particular type of pollutant and/or facility. Fourth, the data base development and data analysis relied upon the combination of several software packages and data bases. Could similar procedures be done using additional software packages and/or data bases to expand into a multimedia effort? These are only a few of the potential future efforts which could be undertaken. Any suggestions of potential future directions and/or methods to achieve them would be appreciated.

References

SAS Institute Inc. Statistical Analysis System (SAS): Version 8E. 2001

EPA office of Pollution Prevention and Toxics. Users Manual for RSEI. Version 2.1. 1998 - 2000

TRI Data. December 2002.

United States Bureau of the Census. 2003 SF3 Detailed Census Tables. Population and Housing SF3 Tables. 2003.